METHODS FOR BIAS REDUCTION IN EVIDENCE-BASED MEDICINE

Arielle K. Marks-Anglin

A DISSERTATION

in

Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

_____

Yong Chen, Associate Professor of Biostatistics, University of Pennsylvania

Graduate Group Chairperson

_____

Nandita Mitra, Professor of Biostatistics, University of Pennsylvania

Dissertation Committee

Rebecca Hubbard, Professor of Biostatistics, University of Pennsylvania

Wei-Ting Hwang, Professor of Biostatistics, University of Pennsylvania

Weijie Su, Assistant Professor of Statistics, University of Pennsylvania

Said Ibrahim, Senior Associate Dean for Diversity and Inclusion, Professor of Healthcare Policy and Research, Professor of Population Health Sciences, Weill Cornell Medical College

METHODS FOR BIAS REDUCTION IN EVIDENCE-BASED MEDICINE

ACKNOWLEDGEMENT

I would like to thank my advisor, Yong Chen, for his tremendous support and mentorship in our four years of working together. His guidance and example have shaped both my work and my perspective as a statistician, encouraging me to see the bigger picture in all my research, and instilling the importance of effectively communicating the relevance of my work. He has been my advocate every step of the way, and I cannot thank him enough for his friendship and encouragement during the best and most difficult stages of this program.

I am also grateful to my committee chair and members: Rebecca Hubbard, Wei-Ting Hwang, Weijie Su and Said Ibrahim, for their dedicated time and effort towards improving the quality of my work. Their insightful questions and recommendations at each committee meeting have proved invaluable for my research. I am especially grateful to Rebecca for always being willing to offer her time, expertise and guidance on working with real world data, as well as providing access to datasets for my research. I am also incredibly thankful for Wei-Ting, who served as my masters thesis advisor and devoted much of her time over the years towards mentoring and collaborating with me.

My work would not be complete with out the invaluable help of my fellow students and postdocs in Yong's PennCIL lab, namely Chongliang Luo, Rui Duan, Mackenzie Edmonson, and Jiayi Tong. Whenever I felt stuck or uncertain, they never hesitated to help and offer advice. I am incredibly grateful to Chongliang in particular, without whose guidance and assistance this work would not have progressed to completion. I am also thankful for my fellow GGEB students, whose friendship made this an enjoyable journey.

I would not be where I am today without the love and support of my family. Mom and Dad, you gave me everything, and my success is because of the values you instilled in me and your hard work that enabled Gianna, Brandon and I to pursue our dreams. Gianna and Brandon, thank you for the love and encouragement you always show me. Finally, I want to thank my core support group and the motivation behind my success: Lawrence and Amayah. Amayah, you inspire me every day to be the best mom and role model I can be for you. I hope that my work will encourage you to pursue your highest goals in life, wherever it may lead. Lawrence, you have believed in me and supported my ambitions and endeavors since day one. Thank you for walking this path with me. May we continue to achieve our dreams together.

ABSTRACT

METHODS FOR BIAS REDUCTION IN EVIDENCE-BASED MEDICINE

Arielle K. Marks-Anglin

Yong Chen

Evidence-based medicine (EBM) emerged as a movement to ground clinical practice in empirical research to optimize patient care and outcomes. The exponential growth in clinical studies that ensued along with the adoption of electronic health records (EHRs) created a cycle of evidence generation, synthesis, translation, and data collection that continues to guide standard of care. The success of EBM hinges on the reproducibility and validity of the research produced. However, systemic bias at any stage can lead to incorrect inference, negatively impacting patient care. In this dissertation, we explore three sources of bias that can undermine EBM, including publication bias in meta-analyses (evidence synthesis), differential outcome misclassification in EHR data (impacting evidence generation), and selection bias in EHR-based studies (evidence translation). For publication bias, we develop an EM-algorithm for selection model estimation in the expanded network meta-analysis (NMA) framework. We show that it substantially reduces bias due to selective publication, while allowing for a maximally flexible working model for heterogeneous data. We apply it to an NMA of antiplatlet therapies for preventing vascular occlusion. For differential misclassification, we propose two surrogate-assisted sampling schemes for cost-effective validation of EHR outcomes. The sampling weights prioritize selection of patients most informative for the model of interest, leading to improved precision of model estimates relative to simple random sampling under measurement constraints. We study their performance under multiple data distributions and offer recommendations for the optimal application of each weighting scheme. We apply our methods to the study of second breast cancer events among women diagnosed with primary stages I-IIIB invasive breast cancer. Finally, we expand the framework of outcome validation to account for patient selection from target populations into EHR cohorts. Combining our efficient sampling designs with inverse probability of selection weighting, we improve the generalizability of results derived from validated subsamples of EHR data. We study a variety of mechanisms for patient selection and the bias-variance tradeoff when constructing sampling weights that account for selection bias. We then use our methods to extend inference from a colon cancer recurrence EHR dataset to the larger

U.S. population diagnosed with stages I-IIIA colon cancer.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

INTRODUCTION: EVIDENCE-BASED MEDICINE (EBM) AND THE EVIDENCE ECOSYSTEM

1.1. Background

David Sackett, regarded by many as the 'father' of evidence-based medicine (EBM), defined EBM as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients," (Sackett et al., 1996). Its practice integrates clinical expertise with the best available evidence along with patient preferences in order to make optimal decisions for care and treatment. The paradigm shift from relying solely on clinical experience, intuition and disease pathophysiology towards an empirically grounded approach to clinical teaching and practice (Guyatt et al., 1992; Pope, 2003) coincided with an exponential growth of published clinical research to unprecedented levels (Alper et al., 2004), and the development and adoption of electronic health record (EHR) systems to improve patient care (Kemper, Uren, and Clark, 2006). Together, these have encouraged a more active approach to EBM, with the emergence of evidence ecosystem (or evidence cycle) models (Cartabellotta and Tilson, 2019; Embi and Payne, 2013; Vandvik and Brandt, 2020) that aim to optimize patient care and outcomes.

Proposals for evidence cycles, also known as learning health systems, take the general form shown in Figure 1.1. At the top of the cycle, evidence is generated through clinical studies (typically randomized controlled trials (RCTs) and observational studies). Due to the vast number of clinical studies available on the comparative benefits/harms of treatments and interventions, the results are often first synthesized in systematic reviews and meta-analyses (MAs) (Berlin and Golub, 2014), which can offer powerful and robust summaries of the evidence-base to inform practice. As the new findings are used to update policy and protocols, and new treatment pathways are implemented in clinical settings, patient-level data is collected and inputted into EHRs, which are then used to generate new evidence and recruit cohorts for future studies.

The success of this ecosystem in improving patient care requires that the evidence generated is reproducible, unbiased and validly translated in order to minimize errors and optimize patient outcomes. However, bias can be introduced at any stage in this system. For example, published

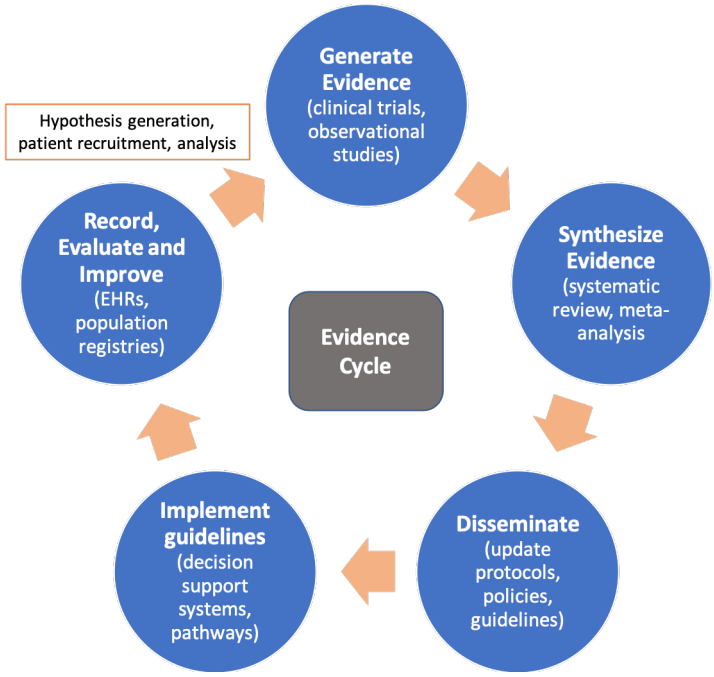


Figure 1.1: The Evidence Cycle

results may be a selective sample of all studies conducted in an area of research, or the patients in an observational study may not be representative of the target population, impacting the success of evidence translation. Finally, quality issues resulting from error-prone EHR data-collection processes can negatively impact the reproducibility of results. This dissertation seeks to explore these sources of bias (specifically at the stages of evidence generation, synthesis, and data collection) and offer solutions to improve evidence-based decision-making.

Systematic reviews and MAs summarize the evidence from a body of research, usually published results from RCTs, to improve the precision of treatment effect estimates and study the consistency of study results. Network Meta-Analysis (NMA), which simultaneously compares multiple treatments and draws on indirect comparisons to strengthen the evidence-base, is particularly useful for making new comparisons between treatments and ranking interventions. MAs and NMAs are generally considered informative for developing treatment recommendations and even operationalizing clinical inputs for cost-effectiveness analyses that impact policy (Khoo et al., 2015). The Cochrane Database is a rich resource that clinicians often refer to to inform their practice. While individual RCTs can generate unbiased causal estimates of treatment effects by the principle of randomization (see Figure 1.2), the larger body of published RCTs is a non-random sample of all

studies conducted, and publication may be dependent on the magnitude and direction of results (see Figure 1.3). This phenomenon is known as 'publication bias' and can induce bias in the summary treatment effect in MAs or the ranking of interventions in NMAs, which in turn affect clinical decisions. It is a known threat to the reproducibility of scientific research (Johnson et al., 2017).

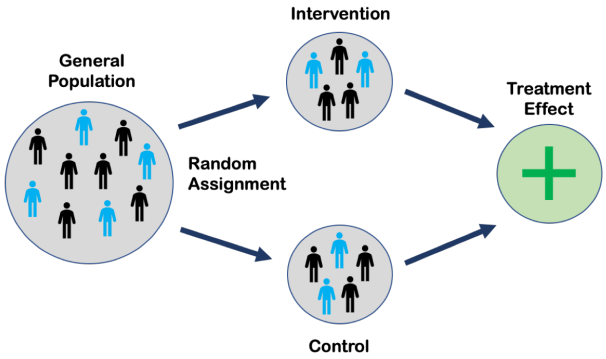


Figure 1.2: Randomized Controlled Trial

On the evidence generation side of the evidence ecosystem, EHR data collected in the process of healthcare delivery is an increasingly used resource in research. For example, large EHR databases are used for studying interventions for rare diseases, identifying cohorts for pragmatic RCTs which aim to evaluate interventions in "real world" settings (Richesson et al., 2013), as well as observational association studies that aid in identifying risk factors for adverse events and building prediction models for individual risk assessment. However, the process of collecting and inputing patient data into electronic records is error prone, and phenotyping algorithms may be used to process patient records and physician notes (see Figure 1.4). Systematic errors in the phenotyping

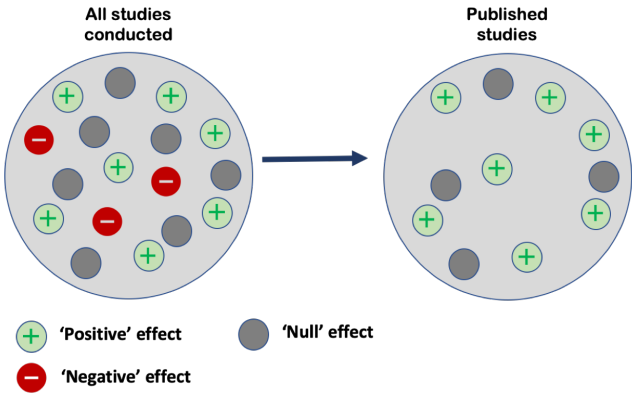


Figure 1.3: Selective Publishing

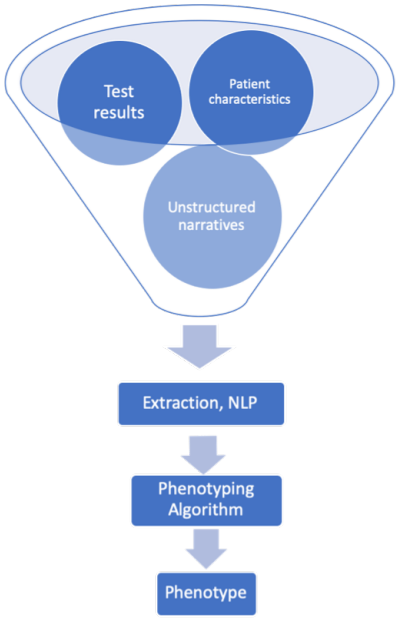


Figure 1.4: Computable Phenotypes

of patient characteristics and outcomes have the potential to severely impact the reproducibility of EHR-based studies, by introducing bias and inflating type I errors. The large size of EHR datasets also means that biased results are estimated with high precision and further compound incorrect inference (Kaplan, Chambers, and Glasgow, 2014). Chart review is generally the gold-standard used for validating misclassified outcomes in EHR data, but is both time consuming and cost-prohibitive. Cost-efficient study designs for outcome validation can be useful for ensuring internally valid estimates from EHR-based studies.

Finally, we consider the transportability of results from EHR-based cohort studies to larger or different clinical populations. EHR databases are not random subsets of the greater population, but rather selective samples developed through patient interactions with specific healthcare systems. This process is influenced by many factors, including (but not limited to) geographic location, health insurance coverage, severity of disease, employment and socioeconomic status. For example, the increasing use of electronic and mobile health data in observational studies and for patient recruitment has caused a 'digital divide' (Ibrahim, Charlson, and Neill, 2020) in inclusion between those with and without access to technology. Additionally, smaller clinics serving rural, uninsured and other underserved groups have lower uptake of EHR systems (Hamamura, Withy, and Hughes,

2017; Hing and Burt, 2009; Mack et al., 2016), and EHR-based studies are more frequently conducted by large, well-resourced, multi-site networks. As a result, the study cohort in an EHR-based study may not be representative of the target population of interest with respect to certain characteristics. For example, EHR databases based in academic institutions underrepresent low healthcare utilizers, including the uninsured and those with limited access to healthcare services (Tikkanen et al., 2017). Failure to account for such selection mechanisms can lead to biased inference and erroneous decision-making.

1.2. Aims

In this dissertation work, we propose the following aims:

Aim 1: Develop a new frequentist approach to selection model estimation for correcting for publication bias in network meta-analysis.

We propose to develop an Expectation-Maximization (EM) algorithm for estimation of a flexible Copas-like selection model to address publication bias in NMAs. We will also evaluate methods for variance estimation under missing reported within-study correlations. The proposed algorithm will be validated empirically through simulation and applied to a network meta-analysis of antiplatlet therapies for maintaining vascular patency (preventing occlusion)

Aim 2: Develop surrogate-assisted sampling schemes for cost-effective chart review of electronic health record data.

We will develop informative sampling weights to guide outcome validation in the logistic regression setting. This method will leverage information from surrogate or misclassified outcomes to improve statistical efficiency as compared to uniform random sampling, or sampling with only the use of covariate data. The method will be evaluated and validated in the setting of differential misclassification of outcomes for EHR data. The proposed method will also be applied to, and validated by, an investigation quantifying risk factors for breast cancer recurrence using an EHR dataset

Aim 3: Address external validity of EHR-based studies that include validation of surrogate EHR-derived endpoints.

We seek to modify the informative sampling approaches from Aim 2 to address the external validity of the results from an EHR-based study that involves validation of surrogate outcomes. Expanding the study framework to make inference on a target population, the proposed methods will balance the joint goals of efficient study design and generalizability by additionally accounting for selection

bias. The sampling scheme will be evaluated through simulation and applied to a colon cancer recurrence dataset from an EHR cohort.

CHAPTER 2

AIM 1: BIAS IN SYSTEMATIC REVIEWS/MAS - DEVELOP A NEW FREQUENTIST APPROACH TO SELECTION MODEL ESTIMATION FOR CORRECTING FOR PUBLICATION BIAS IN NETWORK META-ANALYSIS

2.1. Background

As more clinical research is published, systematic reviews and meta-analyses (MAs) have become increasingly valuable for evaluating and efficiently synthesizing large amounts of evidence on clinical practice and policy. MAs in particular improve the power and precision of estimated effects (Cohn and Becker, 2003), earning a place at the top of the evidence pyramid for informing treatment protocols, policies and guidelines. When evaluating the relative performance of multiple interventions, or direct evidence is limited between new and established treatments, network meta-analysis (NMA) offers advantages over traditional pairwise MA. By aggregating direct and indirect evidence (linear combinations of contrasts with a common comparator) from multiple study designs, NMA provides a framework for simultaneous comparisons between multiple interventions that have not been directly compared, and the inclusion of potentially far more studies than standard MAs (Lu and Ades, 2004; Lumley, 2002). However, the validity of this approach hinges on assumptions of homogeneity and evidence consistency (Song et al., 2003), to which publication bias is a well known threat (Thornton and Lee, 2000). When the dissemination of study findings is dependent on the size and direction of results, systemic heterogeneity is introduced, generating biased results in the relative efficacy and safety of drugs.

Early cohort studies offered empirical evidence of publication bias (Bardy, 1998; Cooper, DeNeve, and Charlton, 1997; Dickersin and Min, 1993), revealing that trials with statistically significant or "positive" findings were more likely to be published or submitted for publishing than those showing non-significant or "negative" results. Such biases can not only lead to inflation of pooled effect estimates in meta-analysis, but also create an incomplete knowledge-base on which to make medical decisions and policies. In an effort to increase transparency and reduce publication bias, the International Committee of Medical Journal Editors (ICMJE) announced in 2004 a requirement for prospective registration of clinical trials as a pre-requisite for consideration for publication in its

member journals, beginning in 2005 (De Angelis et al., 2005). Following this, the U.S. Food and Drug Administration Amendments Act required prospective trial registration with ClinicalTrials.gov for drugs, biologics, and devices subject to Food and Drug Administration (FDA) regulation beginning in 2007 and expanded in 2016 (FDA, 2007; Zarin et al., 2016). The National Institutes of Health (NIH) has also instituted mandated clinical trial registration (Hudson et al., 2016; NIH, 2016). Despite these efforts, not all trials are registered. A meta-analysis recently found that the proportion of registered randomized clinical trials since 2005 was only 53% (Trinquart, Dunn, and Bourgeois, 2018). More importantly, registration of trials does not always lead to publication. Jones et al. (2013) reported that, of 585 large scale ($>$ 500 participants) registered trials, 29% remained unpublished. Similarly, in a recent analysis of registered randomized clinical trials involving digital health interventions, it was found that 27% of trials were unpublished (Al-Durra et al., 2018). Furthermore, while the existing measures have improved the reliability of meta-analyses of more recent studies, those which include older studies continue to suffer from higher levels of selective publishing (Kicinski, Springate, and Kontopantelis, 2015). Publication bias therefore appears to continue to be a concern despite efforts to address the problem.

Substantial work has been done to handle publication bias in the pairwise MA setting, including graph-based methods and selection models (Jin, Zhou, and He, 2015; Marks-Anglin and Chen, 2020a). Graph-based tests (Begg and Mazumdar, 1994; Egger et al., 1997; Schwarzer, Antes, and Schumacher, 2007) focus on evaluating the asymmetry of scatter plots like Galbraith's radial plot (Galbraith, 1988) or the widely used funnel plot (Light and Pillemer, 1984), under the assumption that deviation from a symmetric pattern of estimated treatment effects mapped against some measure of precision is indicative of publication bias. The Trim-and-Fill method (Duval and Tweedie, 2000) was developed as a means of correcting for bias by nonparametrically estimating the number of unpublished studies based on funnel plot asymmetry. Though often recommended as standard practice in univariate MAs, these methods do not readily translate to the network setting, due to the challenges with visualizing asymmetry for direct and indirect contrasts simultaneously . Furthermore, the sole reliance of graphical methods on the symmetry assumption to account for publication bias is controversial, as clinical heterogeneity, choice of outcome measure and even chance can all give rise to asymmetry (Lau et al., 2006; Papageorgiou et al., 2015; Sterne et al., 2011; Terrin, Schmid, and Lau, 2005).

Selection models were proposed as a means of more explicitly characterizing the missing-not-at-random (MNAR) process underlying publication bias. Lane and Dunlap (1978), Hedges (1984, 1992) and Iyengar and Greenhouse (1988) imposed parametric weight functions to the probability of study publication based on p-values or test statistics. Copas and colleagues (Copas and Shi, 2001; Copas and Shi, 2000) later developed a more flexible framework, which relates study publication to the estimated effect size and precision through separate parameters, by specifying a random effects outcome model and separate selection model. Though intuitive, the lack of information from missing studies means that maximizing the complex likelihood using only the observed data often leads to non-identifiability of the selection model parameters, as observed by Carpenter et al. (2009). Copas and colleagues instead proposed a sensitivity analysis using the profile likelihood over a range of selection parameter values. To overcome the limitations of direct inference, Ning, Chen, and Piao (2017) developed an EM algorithm for obtaining a bias-corrected maximum likelihood estimate of a Copas-like model in the univariate setting, imputing missing studies based on both the Copas-based probability of selection and the symmetry property of funnel plots. Unlike the trim-and-fill approach, which imputes missing studies solely based on the symmetry principle, this Copas-based approach is less sensitive to outliers, which can lead to inflated standard errors and more conservative adjustment in trim-and-fill (Schwarzer, Carpenter, and Rücker, 2010).

In comparison to univariate MAs, little has been done in the way of accounting for publication bias in the network setting. Copas' model has only very recently been extended to NMAs, in large part due to the complex correlation structure that needs to be accounted for and substantially increased number of selection parameters for multiple study designs. Chootrakool, Shi, and Yue (2011) were the first to attempt this, specifying a multivariate normal approximation for the multiple treatment comparisons and a separate selection model for each study. Like Copas and Shi (2000), this approach involves maximizing a complex likelihood and is subject to non-convergence issues. To help with identifiability, Mavridis et al. (2014) provided a full Bayesian approach for estimation in the NMA setting. However this requires specification of appropriate prior distributions for several of the model parameters and plausible ranges for the probability of publication, which can be cumbersome for clinical investigators or for which expert opinion may not be available. Furthermore, existing formulations of Copas' model in NMA all assume constant between-study heterogeneity across study designs (known as the 'common variance' assumption), in order to limit the number of parameters that need to be estimated. In doing so, strength is borrowed across comparisons

for heterogeneity assumption. However, this is a strong assumption that may not hold in practice. While violation of this assumption will only impact variance estimates in a standard NMA (Thorlund, Thabane, and Mills, 2013), we show through simulation that it has the potential to severely bias treatment effects in the selection model setting. This is because Copas' model differentiates between random heterogeneity and systemic heterogeneity caused by selective publication when correcting for bias. Misspecification of the random component thus directly impacts to what degree the treatment effect estimates are adjusted.

We propose a frequentist approach to estimation of Copas' model in the network setting that allows for non-constant between-study heterogeneity. As a multivariate extension to Ning, Chen, and Piao (2017), this EM-based Bias Reduction Approach through Copas-model Estimation (EMBRACE) algorithm serves as an alternative to the full Bayesian approach and does not require prior knowledge on the missing studies. And unlike purely graph-based methods, EMBRACE imposes parameterization for the MNAR process, using a multivariate characterization of symmetry only to improve identifiability and numerical stability. By developing an EM algorithm that is devised to be computationally stable for a large number of model parameters, even with the limited samples in NMAs, we can both quantify the non-publication rate and achieve bias reduction for NMAs with two-arm and multi-arm trials. We show through simulation studies that in the presence of publication bias, EMBRACE leads to substantial bias reduction of the effect parameters compared to naïve estimation under the observed likelihood. We also show the importance of correctly specifying the heterogeneity structure. We then apply this method to a network of studies investigating the efficacy of antiplatlet therapies for maintaining vascular patency (unobstructed blood flow) in patients at risk of vascular occlusion. This network has previously been used to illustrate the application of Copas' model for sensitivity analysis (Chootrakool, Shi, and Yue, 2011; Mavridis et al., 2014), though under the common heterogeneity assumption.

This project is organized as follows. Section 2.2 presents the proposed method and derives the EM algorithm. We present simulation studies to evaluate the performance of the proposed method in Section 2.4. In Section 2.5 we apply EMBRACE to a network of antiplatlet therapies and compare the results of our method to previous application of Copas' model. A brief discussion is provided in Section 2.6.

2.2. Method

2.2.1. The network meta-analysis framework

Following the notation presented by Mavridis et al. (2014) for a full NMA setting, we consider a network of n randomized trials comparing a set of T=$\{A, B, C, ...\}$ treatments or interventions. Each study compares a subset of at least two treatments in T, and we denote the subset of treatments compared in study design d as T_d, where $d = 1, 2, \ldots, D$. Thus a given network contains n_d studies of design d, and the D subsets of studies are pairwise disjoint, such that $\sum_{d=1}^{D} n_d = n$.

We denote the estimated effect size for comparing treatments X and Y in the i^{th} study of design d as y_{id}^{XY}, assumed to be approximately normally distributed and centered on the true effect μ^{XY}. These can include continuous outcomes such as differences in means or log-transformed effect measures for binary data. Let $\boldsymbol{y}_{id}$ be the general notation for the observed effect(s) in a given study (scalar for a two-arm study and vector for multi-arm studies with multiple contrasts), and similarly let $\boldsymbol{s}_{id}$. The number of possible contrasts among T_d treatments is calculated as the number of pairs that can be selected without replacement from the set T_d, which is $C_2^{T_d}$. However in practice, only $T_d - 1$ contrasts need to be estimated, as the rest can be obtained through linear combinations. This holds under the assumption of consistency of treatment effects among a collection of studies, in which both direct estimates from head to head trials, and indirect estimates (calculated through a linear combination of estimates from studies which directly compare the treatments of interest against a common comparator), can contribute to inference on a given treatment effect (Salanti et al., 2008). For example, given the summary treatment effects of A vs. C and B vs. C, denoted $\widehat{\mu}^{AC}$ and $\widehat{\mu}^{BC}$, the summary effect estimate of B vs. A, $\widehat{\mu}^{BA}$, can be estimated indirectly as $\widehat{\mu}^{BC}$ - $\widehat{\mu}^{AC}$ if the evidence consistency assumption holds.

2.2.2. Measurement Model

To convey our main idea in its simplest form, we consider an NMA comparing three treatments of interest, $\{A, B, C\}$, although the framework can be extended to larger sets. The published studies are classified into four designs: $T_1 = \{A, C\}$, $T_2 = \{B, C\}$, $T_3 = \{A, B\}$ and $T_4 = \{A, B, C\}$. Thus there are n_4 independent three-arm trials, and the remaining studies (n_1, n_2, n_3) are independent two-arm trials. We wish to estimate μ^{AC}, μ^{BC}, and μ^{BA}, but only need to directly estimate two contrasts, e.g., μ^{AC} and μ^{BC}, under the evidence consistency assumption.

Assuming a random effects model for each outcome, $\boldsymbol{y}_{id} = y_{id}^{XY}$, from the two-arm studies, we have

$$y_{i1}^{AC} \sim N(\theta_{i1}^{AC}, (s_{i1}^{AC})^2), \qquad \theta_{i1}^{AC} \sim N(\mu^{AC}, \tau_1^2) \tag{2.1}$$

$$y_{i2}^{BC} \sim N(\theta_{i2}^{BC}, (s_{i2}^{BC})^2), \qquad \theta_{i2}^{BC} \sim N(\mu^{BC}, \tau_2^2)$$

$$y_{i3}^{BA} \sim N(\theta_{i3}^{BA}, (s_{i3}^{BA})^2), \qquad \theta_{i3}^{BA} \sim N(\mu^{BC} - \mu^{AC}, \tau_3^2)$$

where θ_{id}^{XY} and s_{id}^{XY} denote the study-specific effect size and observed standard error respectively. We assume normality of the study-specific effects, centered at the true underlying effect size, μ^{XY}, with $\{\tau_1^2, \tau_2^2, \tau_3^2, \tau_4^2\}$ representing design-specific between-study heterogeneities. The specification of non-constant heterogeneity is unique to our application of Copas' model, as other formulations (Chootrakool, Shi, and Yue, 2011; Mavridis et al., 2014) typically assume τ^2 is constant across designs.

In the three-arm design, we similarly need only estimate $\{\mu^{AC}, \mu^{BC}\}$. To account for within-study correlation, ρ_{wi}, we specify a multivariate normal distribution with random effects as follows

$$\begin{pmatrix} y_i^{AC} \\ y_i^{BC} \end{pmatrix} \sim N\left(\begin{pmatrix} \theta_i^{AC} \\ \theta_i^{BC} \end{pmatrix}, \begin{pmatrix} (s_i^{AC})^2 & \rho_{wi} s_i^{AC} s_i^{BC} \\ \rho_{wi} s_i^{AC} s_i^{BC} & (s_i^{BC})^2 \end{pmatrix}\right) \tag{2.2}$$

where

$$\begin{pmatrix} \theta_i^{AC} \\ \theta_i^{BC} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu^{AC} \\ \mu^{BC} \end{pmatrix}, \begin{pmatrix} \tau_4^2 & \tau_4^2/2 \\ \tau_4^2/2 & \tau_4^2 \end{pmatrix}\right)$$

2.2.3. Selection model

Copas and Shi (2000) proposed a separate selection model for the publication process, which Mavridis et al. (2014) extended to a treatment-by-design model tailored to the NMA setting. Following their notations, we allow each study of design d to have its own underlying design-specific continuous latent variable, z_{id}, to model the publication process, whereby study i is published ($\boldsymbol{y}_{id}$ is observed) if and only if $z_{id} > 0$. This latent variable can be modeled with design-specific parameters α_d and β_d as

$$z_{id} = \alpha_d + \beta_d / f(\boldsymbol{s}_{id}) + \epsilon_{id} \tag{2.3}$$

where $\epsilon_{id} \sim N(0,1)$, $\beta_d \geq 0$, corr$(z_{id}, \boldsymbol{y}_{id}) = \rho_d$ and corr$(\epsilon_{id}, \boldsymbol{\theta}_{id}) = 0$.

Here $f(\boldsymbol{s}_{id})$ is a pre-specified function of the observed variance-covariance matrix (for example, the standard error in a two-arm trial, or an average for a three-arm trial) conditional on study publication. We differ from Copas and Shi (2000) in assuming the observed within-study variances approximate the true within-study variances in the outcome model (2.2), thus limiting the number of unknown parameters, and refer to model (2.3) as a 'copas-like' model for NMA. Thus α_d controls the marginal publication rate of a study of design d with infinite generalized variance (i.e. very small sample size), and β_d describes how the propensity for publication depends on study variance. β_d is generally assumed to be positive, with larger trials having a greater likelihood of publication than smaller trials. The parameter ρ_d measures the relationship between the observed effect sizes and latent variable z_{id}. Therefore $\rho_d = 0$ implies no publication bias due to the effect size of study design d, while $|\rho_d| > 0$ suggests that larger estimated effect sizes are associated with greater propensity for publication.

Copas' model is unique among selection models in allowing publication to depend separately on effect size and precision, thus accounting for a greater variety of publication processes, some of which do not lead to bias. For example, if $\beta_d \neq 0$ but $\rho_d = 0$, then selection is only impacted by study precision, which does not lead to bias in the summary effects. On the other hand, $\beta_d = 0$ and $\rho_d \neq 0$ will cause bias in the meta-analysis, as selection is dependent on effect size, even if it is unrelated to precision.

Under this selection model, the probability of the i^{th} study of design d being published is $P(z_{id} > 0) = P(\epsilon_{id} > -\alpha_d - \beta_d / f(\boldsymbol{s}_{id})) = \Phi(\alpha_d + \beta_d / f(\boldsymbol{s}_{id}))$, where Φ is the cumulative distribution function of the standard normal distribution. It should be noted that since the parameters in (2.3) are design-specific, then the number of parameters to estimate increases linearly with the number of study designs.

Thus for the two-arm trials, the joint distribution of z_{id} and y_{id}^{XY} is

$$\begin{pmatrix} y_{id}^{XY} \\ z_{id} \end{pmatrix} \sim N\left(\begin{pmatrix} \mu^{XY} \\ \alpha_d + \beta_d / s_{id}^{XY} \end{pmatrix}, \begin{pmatrix} \tau_d^2 + (s_{id}^{XY})^2 & \rho_d s_{id}^{XY} \\ \rho_d s_{id}^{XY} & 1 \end{pmatrix} \right) \tag{2.4}$$

Similarly for the three-arm trials ($T_4 = \{A, B, C\}$):

$$\begin{pmatrix} y_{i4}^{AC} \\ y_{i4}^{BC} \\ z_{i4} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu^{AC} \\ \mu^{BC} \\ \alpha_4 + \beta_4 / f(s_{i4}^{AB}, s_{i4}^{AC}) \end{pmatrix}, \begin{pmatrix} \tau_4^2 + (s_{i4}^{AC})^2 & \tau_4^2/2 + \rho_{wi} s_{i4}^{AC} s_{i4}^{BC} & \rho_4^{AC} s_{i4}^{AC} \\ & \tau_4^2 + (s_{i4}^{BC})^2 & \rho_4^{BC} s_{i4}^{BC} \\ & & 1 \end{pmatrix} \right) \tag{2.5}$$

The distributions in (2.4) and (2.5) simplify the random effects structures in (2.1) and (2.2), centering on the true effect sizes and combining the between-study ($\boldsymbol{\tau^2} = \{\tau_1^2, \tau_2^2, \tau_3^2, \tau_4^2\}$) and within-study variance components in the variance-covariance matrices. Note that for the multi-arm trials, each contrast has its own correlation parameter with z_{i4}, as we perceive that the comparisons have differential impact on the probability of a study's publication.

2.3. Maximum likelihood and the EM algorithm

To overcome the computational difficulties of directly maximizing the likelihood with respect to parameters in (4) and (5), we propose a more stable EM algorithm. The key idea is to correct for bias due to selective publishing by imputing the unobserved (missing) studies. This idea has been found to be effective in the univariate MA setting for estimating two parameters of interest (mean effect size and between-study heterogeneity) (Ning, Chen, and Piao, 2017). In an NMA setting, however, we have to account for a larger number of model parameters as well as the complex variance-covariance matrices.

The estimable parameters in a network meta-analysis with three treatment arms and four study designs are $\psi = (\mu^{AC}, \mu^{BC}, \tau_1, \tau_2, \tau_3, \tau_4, \rho_1, \rho_2, \rho_3, \rho_4^{AC}, \rho_4^{BC}, \alpha_d, \beta_d,\ d = 1, \ldots, 4)$, with μ^{AC}, μ^{BC} and $\boldsymbol{\tau}$ being of primary interest. For simplicity in our illustration and simulations we treat within-study correlation ρ_{wi} as known and constant across all three-arm studies, though for data applications study-specific ρ_{wi}, which ideally are reported in the published multi-arm studies, must be used. In cases where within-study correlation are unreported in multi-arm studies, we offer two options for overcoming this. First, for continuous outcomes or log-transformed measures for binary data, ρ_{wi}

can be calculated from the observed data as

$$\rho_{wi} = \frac{\text{cov}(y_{i4}^{AC}, y_{i4}^{BC})}{s_{i4}^{AC} s_{i4}^{BC}} = \frac{(s_{i4}^{AC})^2 + (s_{i4}^{BC})^2 - (s_{i4}^{AB})^2}{2 * s_{i4}^{AC} s_{i4}^{BC}}, \tag{2.6}$$

since consistency always holds within a multi-arm study i and therefore,

$$var(y_{i4}^{AB}) = var(y_{i4}^{AC} - y_{i4}^{BC}) = \text{var}(y_{i4}^{AC}) + \text{var}(y_{i4}^{BC}) - 2 * \text{cov}(y_{i4}^{AC}, y_{i4}^{BC}). \tag{2.7}$$

$$\Rightarrow (s_{i4}^{AB})^2 = (s_{i4}^{AC})^2 + (s_{i4}^{BC})^2 - 2 * \text{cov}(y_{i4}^{AC}, y_{i4}^{BC}). \tag{2.8}$$

In the event that standard errors are unreported for one of the outcomes in a three-arm trial (note that at least two are required for use of Copas' model), particularly for continuous outcomes (as standard errors can be derived for binary data using sample sizes and event counts), we recommend using the composite likelihood with $\rho_{wi} = 0$ for parameter estimation followed by the sandwich estimator proposed by Chen, Hong, and Riley (2015) for variance estimation.

Due to publication-bias, we assume that for each observed study i in design d there are an additional $m_{id} \geq 0$ studies not observed. This unknown integer variable follows a geometric distribution conditional on the observed data and parameterized by the probability of the i^{th} study of design d being published. Recall that $\boldsymbol{y}_{id}$ and $\boldsymbol{s}_{id}$ denote the observed effect(s) and standard errors(s) in a given study (scalar for two-arm trials, and vectors for multi-arm trials). Similarly, we denote $\boldsymbol{\mu}_d$ as the true treatment effect(s) we seek to estimate for design d, which may also be scalar or vector valued. Then we assume that $P(m_{id} = m) = (1 - P(z_{id} > 0|\boldsymbol{y}_{id}))^m P(z_{id} > 0|\boldsymbol{y}_{id})$, $m = 0, 1, 2, \ldots$. The observed data we denote as $O = \{O_{id}, i = 1, ..., n_d, d = 1, \ldots, 4\} = \{\boldsymbol{y}_{id}, \boldsymbol{s}_{id}, i = 1, \ldots, n_d, d = 1, \ldots, 4\}$, and the unpublished data corresponding to each observed study as $O_{id}^* = m_{id} \cdot \{\boldsymbol{y}_{id}^*, \boldsymbol{s}_{id}\}$, where according to the symmetry assumption, $\boldsymbol{\mu_d} = \frac{\boldsymbol{y}_{id} + \boldsymbol{y}_{id}^*}{2} \Rightarrow \boldsymbol{y}_{id}^* = 2\boldsymbol{\mu}_d - \boldsymbol{y}_{id} = -(\boldsymbol{y}_{id} - 2\boldsymbol{\mu}_d)$. Following the principle of the EM algorithm, we think of $\{O_{id}, O_{id}^*, i = 1, \cdots, N, d = 1, .., 4\}$ as the "complete data" including the observed (published) studies and the unobserved (unpublished) studies, and derive the corresponding log-likelihood as

$$\log L^*(\psi) = \sum_{d=1}^{4} \sum_{i=1}^{n_d} \left[\log \left\{Pr(z_{id} > 0, \boldsymbol{y}_{id})\right\} + m_{id} \log \left\{Pr(z_{id} < 0, -(\boldsymbol{y}_{id} - 2\boldsymbol{\mu}_d))\right\}\right], \tag{2.9}$$

where

$$\begin{aligned}\log\{Pr(z_{id}>0,\mathbf{y}_{id})\} &= \log\{Pr(z_{id}>0|\mathbf{y}_{id})\}+\log\{Pr(\mathbf{y}_{id})\}\\ &\propto \log\{\Phi(w_{id}|\psi)\}-\frac{1}{2}(\mathbf{y}_{id}-\boldsymbol{\mu}_d)^T\text{var}(\boldsymbol{y}_{id})^{-1}(\mathbf{y}_{id}-\boldsymbol{\mu}_d)-\frac{1}{2}\log\{|Var(\boldsymbol{y}_{id})|\}.\end{aligned}$$

By deriving the conditional distribution of $z_{id}|\boldsymbol{y}_{id}$, it can be shown that

$$w_{id}=\frac{\alpha_d+\beta_d/f(\boldsymbol{s}_{id})\ +\ \text{cov}(\boldsymbol{y}_{id},z_{id})^T\text{var}(\boldsymbol{y}_{id})^{-1}(\boldsymbol{y}_{id}-\boldsymbol{\mu}_d)}{\left\{1-\text{cov}(\boldsymbol{y}_{id},z_{id})^T\text{var}(\boldsymbol{y}_{id})^{-1}\text{cov}(\boldsymbol{y}_{id},z_{id})\right\}^{1/2}},$$

where for the two-arm designs, $\text{var}(\boldsymbol{y}_{id})=\tau_d^2+(s_{id}^{XY})^2$, and $\text{cov}(\boldsymbol{y}_{id},z_{id})=\rho_d s_{id}^{XY}$. For the three-arm design,

$$\text{cov}(\boldsymbol{y}_{i4},z_{i4})=\begin{pmatrix}\rho_4^{AC}s_{i4}^{AC}\\ \rho_4^{BC}s_{i4}^{BC}\end{pmatrix},\ \text{var}(\boldsymbol{y}_{i4})=\begin{pmatrix}\tau_4^2+(s_{i4}^{AC})^2 & \tau_4^2/2+\rho_{wi}s_{i4}^{AC}s_{i4}^{BC}\\ \tau_4^2/2+\rho_{wi}s_{i4}^{BC}s_{i4}^{AC} & \tau_4^2+(s_{i4}^{BC})^2\end{pmatrix}$$

are the partitioned components of the variance-covariance matrix in (2.5).

For the expectation (E) step in the EM algorithm, we can calculate the expected value of m_{id} based on the geometric distribution given the observed data and the current parameter estimates (ψ^*) as

$$E[m_{id}|O,\psi^*]=\frac{1-P(z_{id}>0|O,\psi^*)}{P(z_{id}>0|O,\psi^*)}=\frac{1-\Phi(w_{id}|\psi^*)}{\Phi(w_{id}|\psi^*)}$$

Then the expected complete-data log-likelihood function is

$$\begin{aligned}\log L^*(\psi|O,\psi^*)=\sum_{d=1}^{4}\sum_{i=1}^{n_d}[&\log\{Pr(z_{id}>0,\boldsymbol{y}_{id})\}\\ &+\ E[m_{id}|O,\psi^*]\log\{Pr(z_{id}<0,-(\boldsymbol{y}_{id}-2\boldsymbol{\mu}_d))\}]\end{aligned}\tag{2.10}$$

In the maximization (M) step, we maximize the above conditional expected log-likelihood and update the parameter estimates accordingly, iterating between the E and M steps until convergence is achieved according to some pre-specified criterion.

2.3.1. A note on variance estimation

For calculating the observed information for estimates from am EM algorithm, the Louis variance formula is often used (Louis, 1982). It computes the observed information matrix under the complete-data log-likelihood (see equation (2.10)). However, in our algorithm, $E[m_{id}|O, \psi^*]$ has the potential to increase dramatically if $P(z_{id} > 0|O, \psi^*)$ is very small, leading to a very large complete dataset and therefore unreasonably small variance estimates. As the aim of sensitivity analysis for publication bias is bias correction, we propose computing the information matrix under the observed log-likelihood as a working variance-covariance structure. In this case confidence intervals will be similar in size to the naïve estimates, but shifted to reflect the bias correction.

Given the final estimates and covariances of $\widehat{\mu}^{AC}$ and $\widehat{\mu}^{BC}$, by evidence consistency we have

$$\widehat{\mu}^{BA} = \widehat{\mu}^{BC} - \widehat{\mu}^{AC} \tag{2.11}$$

$$\mathrm{var}(\widehat{\mu}^{BA}) = \mathrm{var}(\widehat{\mu}^{BC}) + \mathrm{var}(\widehat{\mu}^{AC}) - 2\mathrm{cov}(\widehat{\mu}^{BC}, \widehat{\mu}^{AC}).$$

2.4. Simulation study

We performed a simulation study to validate the proposed method and EM algorithm, by comparing it to the standard unadjusted (naïve) NMA under a naive log-likelihood. We sought to evaluate the performance and robustness of the algorithm under different rates of selective publication and sample size.

Following the framework described above, we considered a three-treatment network with four study designs: $\{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}$, with true treatment effects $(\mu^{AC}, \mu^{BC}, \mu^{BA}) = (0.5, 0.8, 0.3)$. To evaluate our method in a small-to-moderate sample size setting, we generated $n_d = \{10, 25\}$, $d = 1, 2, 3, 4$, observed study outcomes and variances (for a total of $n = \{40, 100\}$ studies) using the joint models in (2.4) and (2.5). We assumed a constant within-study correlation of $\rho_{wi} = 0.3$ and allowed for design-specific heterogeneity, with $(\tau_1, \tau_2, \tau_3, \tau_4) = (0.4, 0.6, 0.8, 1.0)$. Selection parameters α_d and β_d were chosen such that studies were selected with non-publishing rates (NPR) of 30% and 50%, reflecting moderate to large study selection. Empirical mean bias and relative bias were calculated over 500 simulates for $\widehat{\mu}^{AC}, \widehat{\mu}^{BC}$, $\widehat{\mu}^{BA}$ and $\widehat{\boldsymbol{\tau}}$ for the proposed method and compared with their counterparts from the naïve model assuming no selection bias

($\rho_d^{AC} = \rho_d^{BC} = 0$). Rates of non-convergence were also studied for the proposed algorithm, with non-convergence determined after 1000 iterations. An outline of this simulation study design can be found in Table 1.

Table 2.1: Settings in simulation study design

τ^2	(0.4,0.6,0.8,1.0)			
NPR	0.3		0.5	
n_d	10	25	10	25

We further investigated the performance of Copas' model when common between-study heterogeneity is assumed (i.e. $\tau_1 = \tau_2 = \tau_3 = \tau_4$), as well as the sensitivity of results to initial values used in the EM algorithm.

2.4.1. Performance of EMBRACE compared to naïve NMA

Simulation results are included in Table 2.2, which report findings for the 30% nonpublication and 50% nonpublication settings. In both settings, substantial bias reduction was achieved for all estimated treatment effects relative to the naïve approach, ranging from 69% to 112% . The increase in model-based variance due to the larger number of parameters in the selection model can be quantified at 2%-14% when the non-publishing rate is 30%, and 26%-44% when the non-publishing rate is 50%, with the largest variances corresponding to $\widehat{\mu}^{BA}$, which is indirectly estimated.

Finally, the EM algorithm is shown to be stable when initial starting values for selection model parameters are close to the truth, converging (with precision of 10^{-6}) in 98%-100% of replicates.

2.4.2. Sensitivity of results to mispecification of heterogeneity structure

We also investigated the performance of Copas' model when formulated with fully structured, common between-study heterogeneity (i.e. $\tau_1 = \tau_2 = \tau_3 = \tau_4$), as opposed to the true data generating mechanism which is unstructured, allowing for different levels of heterogeneity by study design. In Table 1, Appendix B we see that misspecifying the heterogeneity structure only slightly inflates the model-based variance of the estimates in a naïve random-effects analysis assuming no selection. This is consistent with the findings of Thorlund, Thabane, and Mills (2013). However, when we attempt to model the selection mechanism, assuming common heterogeneity leads to over-correction for certain treatment effects by the algorithm.

Table 2.2: Empirical bias of parameter estimates using proposed method vs. naïve random effects model; calculated over 500 simulations under non-publishing rates of 30% and 50%, with $n_d = 25, 50$ per study design, $\{\tau_1^2, \tau_2^2, \tau_3^2, \tau_4^2\} = \{0.4, 0.6, 0.8, 1.0\}$, $\mu^{AC} = 0.8$, $\mu^{BC} = 0.5$, and $\mu^{BA} = -0.3$. Bias entries are multiplied by 100.

		Naïve NMA				EMBRACE				
n_d	PAR	MEAN BIAS	ESE	MBSE	REL BIAS	MEAN BIAS	ESE	MBSE	REL BIAS	BIAS RED
non-publishing rate = 30%										
10	μ^{AC}	7.6	18.1	16.3	9.4	2.3	19.7	18.2	2.9	69%
	μ^{BC}	12.4	20.2	17.8	24.9	2.4	22.3	19.2	4.8	81%
	μ^{BA}	4.9	19.5	19.0	-16.3	0.0	21.2	21.6	-0.1	99%
25	μ^{AC}	7.0	11.3	10.6	8.8	1.0	12.2	12.1	1.3	86%
	μ^{BC}	10.9	11.5	11.5	21.9	0.5	13.9	11.7	1.1	95%
	μ^{BA}	3.9	12.4	12.2	-13.0	-0.5	14.3	13.9	1.6	112%
non-publishing rate = 50%										
10	μ^{AC}	9.0	19.1	16.1	11.3	-0.2	21.7	22.1	-0.3	102%
	μ^{BC}	15.7	19.1	17.9	31.4	-0.4	26.8	23.6	-0.8	102%
	μ^{BA}	6.7	21.5	18.9	-22.3	-0.2	27.3	27.3	0.6	103%
25	μ^{AC}	9.9	10.9	10.5	12.4	-0.7	14.4	13.2	-0.9	107%
	μ^{BC}	16.2	12.2	11.5	32.3	-0.1	17.6	14.7	-0.3	101%
	μ^{BA}	6.2	12.6	12.1	-20.7	0.6	17.3	16.3	-1.9	91%

ESE: empirical standard error; MBSE: model based standard error using the first 3×3 submatrix of the observed information matrix; REL BIAS: relative empirical bias; BIAS RED: percent bias reduction relative to naïve estimates
BIAS, ESE, MBSE entries multiplied by 100

Such results may be because under a misspecified formulation, if heterogeneity is underestimated, Copas' model then attributes a greater proportion of between-study variability to selection bias rather than random heterogeneity, resulting in a greater number of imputed studies in the expectation step of the EM algorithm. Alternatively, if heterogeneity is overestimated, Copas' model will underestimate the number of missing studies due to selection bias, leading to undercorrection. Under the misspecified model, we found that the correlation parameters ρ_d for designs 3 and 4 were overestimated, while the heterogeneity estimates of approximately $\widehat{\tau^2} = 0.84^2 = 0.70$ underestimate $\{\tau_3^2, \tau_4^2\} = \{0.8, 1.0\}$, resulting in greater correction by Copas' model.

2.4.3. Sensitivity of results to initial values of EM algorithm

A known issue with the EM algorithm is that successful convergence and attaining of the global maximum may be dependent on the starting values if there are many local maxima (Karlis and Xekalaki, 2003; Laird, 1978). In our simulations, treatment effect estimates were initialized at

their naïve values, while heterogeneity parameters were initialized at the moment estimators using univariate comparisons. Selection parameters ($\rho_d^{AC}, \rho_d^{BC}, \alpha_d, \beta_d$) were initialized close to their true values with some random error ($+\kappa \sim Uniform(-0.1, 0.1)$). As shown in Section 2.4.1, successful bias reduction was achieved using these initial values. However, little information may be available in practice to inform starting values for the selection parameters ($\rho_d, \alpha_d, \beta_d$). We therefore assessed algorithm performance when starting values for the selection parameters are farther from the truth (eg. $+\kappa \sim Uniform(-0.5, 0.5)$ for α_d and β_d, or between 0 and 1 for $\{\rho_1, \rho_2, \rho_3, \rho_4^{AC}, \rho_5^{BC}\}$.

The results in Table 2, Appendix B show poor convergence and high variability of results (evidenced by the empirical standard error) when all selection parameters are initialized farther from their true values. Interestingly, performance appears to be substantially governed by the β_d parameters, as the algorithm has near optimal convergence rates if at least β_d is initialized close to the truth, while the other parameters have starting values that are farther away, though some extra variability in the estimates remains. Stability may be further improved with tuning of initial values for ρ_d, though this does not appear to be essential.

Based on these results, we recommend running the algorithm with multiple sets of initial values for ρ_d^{AC}, ρ_d^{BC}, α_d and β_d, with finer granularity over the range of values for β_d, since this has the greatest impact on outcomes. Specifying the range of values in the grid can be guided using external information where available. Alternatively, note that $\{\rho_d^{AC}, \rho_d^{BC}\}$ are necessarily limited at (-1,1), though we recommend narrowing the range further (eg. (-1,0) or (0,1)) based on the direction of study selection observed in funnel plots. R package `metafor::trimfill` can aid in identifying the side of the plot with greater missingness. A recommended range for α_d is (-2,2) (for a study of infinite variance, these limits correspond to marginal probabilities of selection of 2% and 98% respectively). The bounds on initial values for β_d can be specified using the limits of α_d and the standard error of the most precise observed study, as follows:

$$\beta_{d,lower} = 0, \qquad \text{and} \qquad \beta_{d,upper} = s_{id,\text{min}}(\Phi^{-1}(0.98) + 2)$$

After running the algorithm for each set of initial values, final results can be chosen that maximize the expected complete data log-likelihood.

2.5. Application to Antiplatlet Data

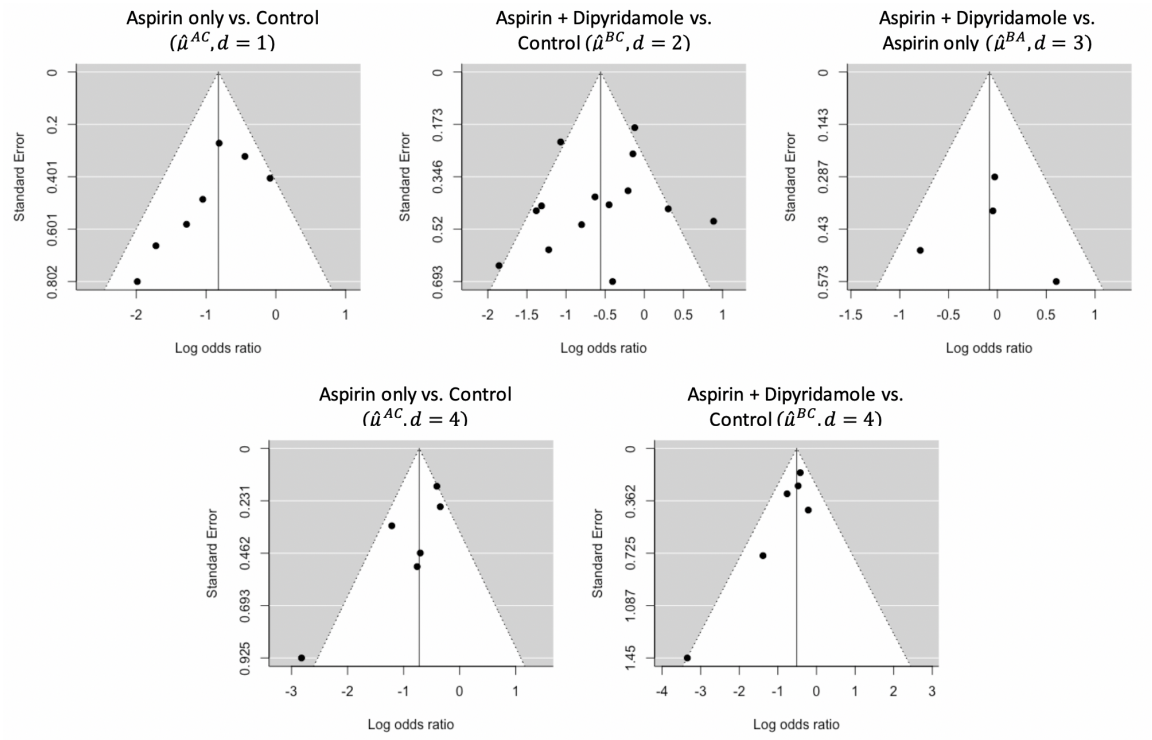
To illustrate the application of our method in practice, we analyzed a network of 31 published RCTs evaluating the efficacy of 3 antiplatlet therapies for maintaining vascular patency (preventing re-occlusion) in patients who have undergone procedures such as bypass grafting or angioplasty to restore blood flow. The interventions compared were Aspirin alone (A), Aspirin + Dipyridamole (B) and a control (C). 7 studies compared A vs. C ($d = 1$), 14 compared B vs. C ($d = 2$), 4 compared A vs. B ($d = 3$) and 6 multi-arm trials compared A vs. B vs. C ($d = 4$).

Figure 2.1 displays funnel plots of the log odds ratios (OR) for two-arm comparisons, including those from multi-arm studies. Patterns of asymmetry are observed in 3 of the 5 plots, with the plot for study design $d = 3$ being difficult to interpret due to the small sample size. Egger's regression test for funnel plot asymmetry using the `metafor` R package yielded significant p-values for Aspirin vs. Control ($\widehat{\mu}_{AC}, d = 1, 4$), and Aspirin + Dipyridamole vs. control in the multi-arm studies($\widehat{\mu}_{BC}, d = 4$). The test for inconsistency in the `netmeta` function in R was used to evaluate inconsistency between study designs, which yielded a p-value 0.57, suggesting the consistency assumption holds in this network.

For the EM algorithm, noting the potential for initial values to determine the maxima, we initialized parameters at varying values and chose final estimates that corresponded to the global maximum. The effect sizes (μ^{AC}, μ^{BC}) were initialized at their naïve NMA estimates under the naive observed log-likelihood, while τ was intialized using the moment estimator. Based on the direction of funnel plot asymmetry, correlations between the outcome and selection models, ($\rho_1, \rho_2, \rho_3, \rho_4^{AC}, \rho_4^{BC}$), were initialized randomly between $\{(-1, 0), (-1, 0), (0, 1), (-1, 0), (-1, 0)\}$ respectively. Starting points for the remaining selection model parameters ranged from -2 and 2 for $\boldsymbol{\alpha_d}$, and 0 to 1 for $\boldsymbol{\beta_d}, d = 1, 2, 3, 4$. Out of several iterations, the final estimates were chosen which maximized the expected log-likelihood. Within-study correlations were derived using equation (2.6). Confidence intervals were calculated under the observed log-likelihood, as discussed in Section 2.3.1.

The final estimates were compared with the naive results as well as those produced using the Bayesian approach by Mavridis et al. (2013). Figure 2.2 shows a forest plot comparing the naïve estimates with those from EMBRACE. The proposed method is shown to have substantially attenuated the log-odds ratio estimates $\widehat{\mu}^{AB}$ (from -0.66 (95% CI: -0.88, -0.44) to -0.41 (95% CI:

Figure 2.1: Funnel plots of estimated log odds ratios and standard errors from published studies on efficacy of antiplatlet therapies in preventing vascular occlusion. Upper panel: contrasts from univariate studies only ($d = 1, 2, 3$). Lower panel: contrasts from multi-arm studies ($d = 4$)

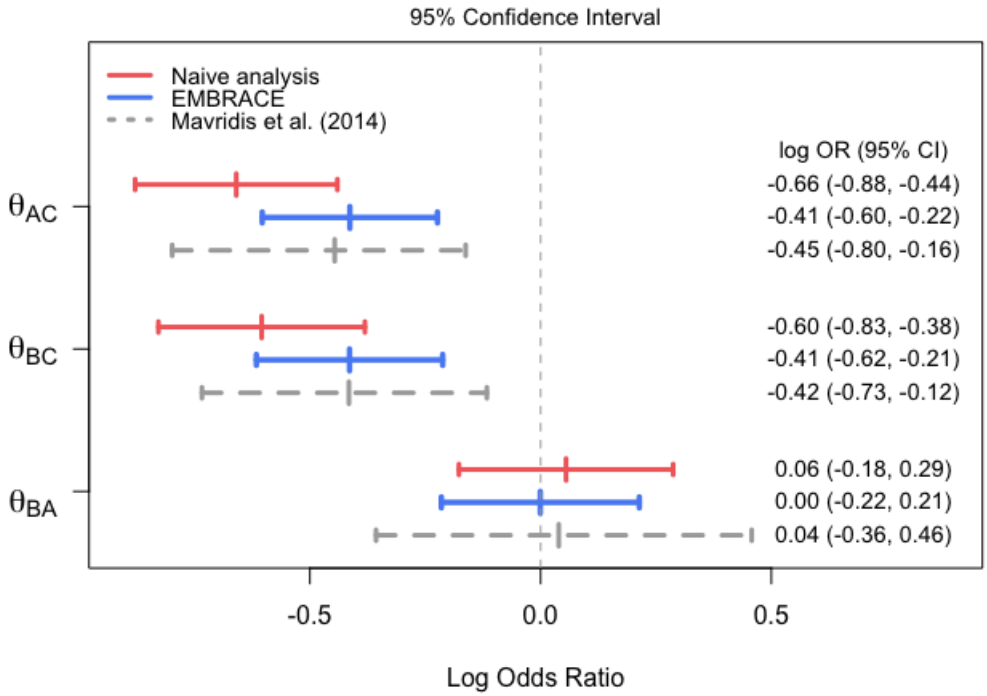


-0.59,-0.23))) and $\widehat{\mu}^{AC}$ (from -0.60 (95% CI: -0.83, -0.38) to -0.41 (95% CI: -0.63, -0.20)), also reducing the difference in effectiveness between B and C to 0 (95% CI: -0.22,0.22). These results align most closely with estimates from scenario 4 in Mavridis et al.'s (2014) Bayesian sensitivity analysis, though they are not equivalent, likely due to the influence of priors on selection parameter estimates in the Bayesian framework and our use of an unstructured heterogeneity model, which was shown in simulations (see Web Appendix A) to result in less biased estimates.

2.6. Discussion

The usefulness of NMAs for evidence generation and decision-making is hindered by the selective publishing of clinical studies, which compromises the underlying assumption of homogeneity and leads to biased results. Due to the computational difficulties associated with this finite sample, MNAR problem, only sensitivity analysis and Bayesian solutions to inference using Copas' model have been developed for NMA. In this paper we have presented a frequentist alternative

Figure 2.2: Forest plot of estimated log odds ratios and confidence intervals from standard NMA of 31 published trials on efficacy of antiplatlet therapies in preventing vascular occlusion, and adjusted estimates when applying EMBRACE to the published studies.



using a novel EM algorithm for a Copas-like selection model in the NMA framework to enable bias quantification and correction. It makes use of the symmetry property of funnel plots to improve identifiability of the model parameters, in contrast to the Bayesian approach by Mavridis et al. (2014) which involves specifying prior distributions for the upper and lower bounds of non-publishing probabilities for each study design. We calibrated our proposed method in simulations and found that our method achieves considerable bias reduction for key parameters of interest relative to the standard unadjusted approach under moderate to severe publication bias, with high levels of algorithmic convergence. We further showed the advantages of allowing for unstructured heterogeneity in bias correction methods, and explored the cost of assuming a common-variance structure.

These results are promising when we consider the ratio of observations to model parameters. The network of treatments in our particular data example contained only 31 published studies for estimation of 19 parameters, resulting in a ratio of 1.6 studies per parameter. The larger datasets

generated in simulation provided at most 5.3 studies per parameter. Of note is that the degree of bias reduction achieved did not vary substantially between small or large sample sizes or rates of non-publication.

We acknowledge that this approach is not without limitations. The sensitivity of the EM algorithm to initial values can be challenging in practice, though we offered recommendations for attaining the global maximum in Section 2.4.3. We further showed that the algorithm is most sensitive to only one of the design-specific selection parameters, β_d, which relates the probability of publication to study precision. Based on our guidelines for starting values, good convergence rates were achieved with appropriate bias reduction. A more general critique of Copas' model is that, as a more sophisticated method compared to graph-based approaches like trim-and-fill and Egger's test, there may be a barrier to use by clinicians with less statistical training (Marks-Anglin and Chen, 2020b). We hope that the release of a general R package currently under development will aid in implementing the proposed algorithm in practice. We also note that asymmetry in funnel plots (which we use to overcome non-identifiability in our model) can arise from sources unrelated to publicaton bias (Sterne et al., 2011). However, in our algorithm the number of missing studies imputed and the distribution of missing outcomes rely entirely on the parameterization of Copas' model. Furthermore, in simulations we generated publication bias according to Copas' model, not asymmetry, yet substantial bias reduction was achieved. Finally, evaluation and correction for publication bias should always be preceeded by an assessment of evidence consistency and transitivity, conditions which need to be met in a valid NMA. However, note that publication bias can itself lead to inconsistency in a treatment network, thus inconsistency may not necessarily preclude an evaluation of selection bias.

Our method employs a maximally flexible working model in the network setting, assuming different selection mechanisms for each study design. While bias correction is the first priority, more robust variance estimates should be explored. Furthermore, application of this model to large networks, where some study designs may only consist of one or a few studies, may require collapsing some or all of the selection model parameters in order to borrow strength across study designs. Such a parsimonious model effectively assumes a common selection mechanism and can further improve computational stability. For example, one may wish to assume a common marginal probability of selection across study designs within a single therapeutic area, which based on our simulations

may be reasonable (see Appendix B). Future work should investigate the robustness of this reduced model to model misspecification, for example when there are in fact different selection mechanisms.

Overall, we believe the proposed method to be an improvement over existing Copas formulations that assume a common heterogeneity structure, as well as graph-based methods like trim-and-fill which do not readily extend to the network setting and may be sensitive to outliers. Furthermore, the EM-algorithm does not require prior knowledge for estimation as with the Bayesian approaches developed by Mavridis et al. (2014), although such knowledge could aid in initializing the parameters for computation. We hope this bias correction method will advance evidence synthesis methods with high methodological rigor.

CHAPTER 3

AIM 2: MISCLASSIFICATION BIAS IN EHR-BASED STUDIES - DEVELOP A SURROGATE-AUGMENTED SAMPLING SCHEME FOR COST-EFFECTIVE CHART REVIEW OF ELECTRONIC HEALTH RECORD DATA

3.1. Background

Electronic health record (EHR) data are increasingly utilized for clinical research due to the tremendous amount of patient data available and the extensive health information contained in them (Hripcsak and Albers, 2012; Jensen, Jensen, and Brunak, 2012), 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 (Forrest et al., 2014)), evaluate treatment effects in diverse, non-trial populations, identify new indications for drug repurposing (Wu et al., 2019) and predict adverse events for drug usage (Menendez, Janssen, and Ring, 2016). In order to utilize EHRs in clinical studies, processing of the raw (including structured and unstructured) 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 (for examples, see Hubbard et al., 2019; Kirby et al., 2016). However, quality issues in EHR data (including inaccurate data and data fragmentation (Wang et al., 2020)) can lead to error-prone phenotyping algorithms, resulting in misclassification or measurement error in the derived variables. The use of misclassified variables in EHR-based studies poses a major threat to the reproducibility (Denaxas et al., 2017) of EHR-based discoveries, and undermines its promise in expanding the horizons of traditional clinical research.

Exposure-dependent phenotyping error for disease status (a form of differential misclassification) is 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 (Duan et al., 2016), inflated type I errors (Chen et al., 2019) and biased association parameter estimates (Hubbard et al., 2020; Neuhaus, 1999). This can occur when certain patient groups have more information in their health record (due to a larger number of visits at the same institution or better documentation of health status) to enable accurate phenotyping. Phenotyp-

ing errors could therefore dramatically affect the results of clinical studies, including comparative effectiveness research where various interventions (e.g., medications, procedures, surgeries) are compared against each other.

Manual chart review remains the gold standard for validating EHR-derived phenotypes and ensuring high quality data for clinical research (Martin et al., 2017). 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 usually only a subset of the sample can be validated. Several methods have been developed for bias reduction and improved efficiency in the presence of outcome misclassification in semi-supervised settings (Chakrabortty and Cai, 2018; Cheng, Ananthakrishnan, and Cai, 2020; Tong et al., 2020), 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 is used to improve precision, sometimes with the aid of surrogate variables. However, these advances in semi-supervised learning focus on a randomly sampled validation set. In the context of EHR data based research, investigators have the opportunity to consider alternative designs for selecting validation sets, which can improve precision of the 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 computer science literature on algorithmic leveraging to propose validation study designs for EHR data that improve statistical efficiency. Algorithmic leveraging is a sampling process that selects a subset of observations according to a probability distribution that depends on empirical statistical leverage (or 'importance') scores to identify the most influential observations for a given model (Ma, Mahoney, and Yu, 2015). It has traditionally been applied to large-scale matrix problems (including least squares approximation (Drineas, Mahoney, and Muthukrishnan, 2006) and low-rank matrix approximation (Mahoney and Drineas, 2009)) 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 analyze the full sample, computing would instead be performed on the smaller selected subsample, and the estimate would approximate that using the full dataset. In recent years, algorithmic leveraging has been proposed to improve statistical efficiency (defined using the

mean squared error) in the setting of linear regression (Ma, Mahoney, and Yu, 2015; Ma and Sun, 2015) and logistic regression (Fithian and Hastie, 2014; Wang, Zhu, and Ma, 2018; Zhang, Ning, and Ruppert, 2019).

For the setting of EHR data with misclassified outcomes, we are interested in developing validation sampling approaches for estimation of a logistic regression model of gold-standard outcome, y, conditional on a feature vector x. Existing methods for efficient sampling for logistic regression either assume that the gold-standard outcome is observed for everyone (Wang, Zhu, and Ma, 2018), or missing entirely (Zhang, Ning, and Ruppert, 2019), 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, which is not accounted for in the weights proposed by Zhang, Ning, and Ruppert (2019). Alternatively, in lieu of a computable phenotype, 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 (eg. the effect sizes of risk factors) would not have their intended interpretations.

In this paper we propose two surrogate assisted sampling schemes that make use of both the covariates, $\boldsymbol{X}$, and the surrogate outcome, s (i.e. the observed, potentially misclassified outcome) for improving statistical efficiency in outcome validation study designs using EHR data. 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. Both proposed sampling designs aim for the resultant estimator based on the subsample to approximate the full data maximum-likelihood estimator (MLE) (the estimate if we had observed the gold-standard outcome for all individuals). This is achieved by applying the A-optimality criterion (minimizing the trace of the matrix) (Chan, 1982) to the asymptotic approximation error. We derive one set of weights along the lines of Zhang, Ning, and Ruppert (2019), which are similarly designed for the measurement-constrained setting, but we incorporate additional information contained in the surrogate variable s. This involves a two-step sampling scheme in which the gold standard, y, is validated through chart review for a small random subsample in the first step. Pilot estimates, along with the surrogate s and covariates $\boldsymbol{X}$, are then incorporated in the second step sampling weights to guide chart review for the remaining study sample. Our second set of proposed weights, which can be implemented in a single step, are A-optimal for the logistic regression model of s regressed

on $\boldsymbol{X}$, which we show still offers information on the influence of observations for the model of interest. We apply these approaches to a real EHR dataset from Kaiser Permanente Washington (KPWA) to study risk factors for second breast cancer events (SBCE) in women with a personal history of breast cancer, and compare costs associated with chart review.

This article is organized as follows. In Section 3.2 we describe the general sampling framework for logistic regression, and introduce the proposed surrogate assisted sampling weights. In Section 3.3 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 3.4 we apply the various weighting approaches to the KPWA study of second breast cancers. A brief discussion is offered in Section 3.5.

3.2. Methods

3.2.1. Setting and Logistic Regression

To illustrate our methods, we denote the full data matrix for n subjects as $\mathscr{F}_n = (\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{s})$, 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 phenotypes or a proxy variable associated with the outcome. We assume that $y_i, s_i \in \{0, 1\}$ and $\boldsymbol{x}_i \in \Re^p$.

Our interest is in estimating the coefficient vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$, which belongs to a compact set in $\Re^p$, from the following logistic regression model of y_i conditional on $\boldsymbol{x}_i$,

$$\operatorname{logit}\{p(y_i = 1|\boldsymbol{x}_i)\} = \boldsymbol{x}_i^T\boldsymbol{\beta}, \tag{3.1}$$

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

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

where $p_i(\boldsymbol{\beta}) = \exp(\boldsymbol{x}_i^T\boldsymbol{\beta}) / \left\{1 + \exp(\boldsymbol{x}_i^T\boldsymbol{\beta})\right\}$. 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 $\boldsymbol{w}_i = (\boldsymbol{x}_i, s_i)$ is observed for the full sample, and y_i must be uncovered through chart review. As it is time-consuming and cost-prohibitive to validate outcomes for the entire dataset, we can only afford to determine y_i for $r \ll n$ individuals. As discussed earlier, Zhang, Ning, and Ruppert (2019) developed a sampling scheme to select an informative subsample of size r on which to obtain y, however they only incorporated information from the covariates $\boldsymbol{x}_i$. Our goal is to develop a sampling framework that makes use of both s_i and $\boldsymbol{x}_i$.

3.2.2. General Sampling Scheme for Outcome Validation in Logistic Regression

In Algorithm 1 we outline a general subsampling algorithm for outcome validation with logistic regression models. 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 $\tilde{\boldsymbol{\beta}}$.

In the setting where the gold-standard outcome is observed for all individuals, Wang, Zhu, and Ma (2018) proved that $\tilde{\boldsymbol{\beta}}$ is consistent for $\hat{\boldsymbol{\beta}}_{MLE}$, conditional on the full data matrix $\mathscr{F}_n$, as $n \to \infty$ and $r \to \infty$ (see section S.1.1 of Wang, Zhu, and Ma (2018)), under certain regularity conditions on the covariate distribution. Furthermore, they show 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 $\mathscr{F}_n$,

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

where

$$\boldsymbol{V} = \boldsymbol{M}_X^{-1}\text{Var}(\psi_i|\mathscr{F})\boldsymbol{M}_X^{-1} = O_p(r^{-1}), \qquad \psi_i = \frac{\left\{y_i^* - p_i^*(\tilde{\boldsymbol{\beta}})\right\}\boldsymbol{x}_i^*}{n\pi_i^*}$$

and $\boldsymbol{M}_{\boldsymbol{x}} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} p_i(\hat{\boldsymbol{\beta}}_{MLE})\left\{1 - p_i(\hat{\boldsymbol{\beta}}_{MLE})\right\}\boldsymbol{x}_i\boldsymbol{x}_i^T$.

Algorithm 1: General subsampling algorithm for outcome validation, adapted from Wang, Zhu, and Ma (2018)

1. Sample $r(\ll n)$ data points with replacement from the full dataset 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 following weighted log-likelihood function to obtain an estimate, $\tilde{\boldsymbol{\beta}}$ based on the subsample. Specifically,

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

   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:

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

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

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

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.

3.2.3. Prior work on optimal sampling weights

A simple choice for the sampling weights in Algorithm 1 is $\pi_i = 1/n$. This is also known as *uniform sampling*. However, this may not be the "optimal" choice in the sense that $\tilde{\boldsymbol{\beta}}$ may be estimated with greater precision under alternative weights. Motivated by large-scale data problems, Wang, Zhu, and Ma (2018) derived optimal weights for the setting where y_i is observed for all individuals. We

denote these weights as $\pi_{i,y\text{OBS}}$ (reflecting that they are optimal when y_i is observed for the full sample),

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

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

They propose a two-step sampling approach, in which an initial sample of size r_1 is selected with uniform probability to obtain a preliminary estimate of $\hat{\boldsymbol{\beta}}_{MLE}$. This enables us to characterize the relationship between y_i and $\boldsymbol{x}_i$ and therefore to determine which observations are 'surprising' given their expected value. The pilot estimate of $\hat{\boldsymbol{\beta}}_{MLE}$ is then included in the calculation of the $\pi_{i,y\text{OBS}}$ for selection of a second, more informative subsample for estimating $\tilde{\boldsymbol{\beta}}$.

For the setting where y_i is unknown and only $\boldsymbol{x}_i$ is observed (measurement constrained setting), Zhang, Ning, and Ruppert (2019) proposed a two-step outcome validation scheme similar to Wang et al.'s work, but in which y_i must be uncovered for individuals selected in the step 1 and step 2 samples. We denote these weights as $\pi_{i,y\text{MISS}}$, reflecting that they are intended for settings where y_i is missing and only obtained for individuals selected in the subsample. The weights are constructed using pilot parameter estimates for the modelled relationship between $\boldsymbol{x}_i$ and y_i (estimated using the step 1 sample),

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

Note that unlike the weights given by Wang, Zhu, and Ma (2018) ($\pi_{i,y\text{OBS}}$), 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, Ning, and Ruppert (2019) 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 $\pi_{i,y\text{MISS}}$ gives greatest weight to individuals for whom the model is most uncertain about their outcome. This approach, which is a form of 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_i to offer additional insight on which individuals will have a y_i value that is 'surprising' given their fitted probabilities $p_i(\hat{\boldsymbol{\beta}}_{MLE})$.

3.2.4. Surrogate-Assisted Sampling for Outcome Validation

Having reviewed the 'optimal' weights for settings 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. 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 $\pi_{i,y\text{MISS}}$).

We propose two sets of candidate weights to use in place of the targeted optimal weights, $\pi_{i,y\text{OBS}}$ (i.e. the optimal weights if $\boldsymbol{y}$ were observed, proposed by Wang, Zhu, and Ma (2018) according to the A-optimality criterion) in the measurement constrained setting.

Surrogate augmented weights

If y_i is known, one can proceed to derive the A-optimal sampling weights, $\pi_{i,y\text{OBS}}$, that minimize the asymptotic approximation error of $(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})$ by minimizing the trace of $\boldsymbol{V}$, as is done in Wang, Zhu, and Ma (2018). 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, Ning, and Ruppert (2019) first applied this formula for a single component, $\boldsymbol{y}$. We propose to additionally apply it for $\boldsymbol{s}$ as follows,

$$
\begin{aligned}
\text{Var}\left\{(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})|\boldsymbol{X}\right\} &= \text{E}\left\{\text{Var}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s},\boldsymbol{y},\boldsymbol{X})|\boldsymbol{s},\boldsymbol{X}\right\} + \\
&\quad + \text{E}\left\{\text{Var}(\text{E}[\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s},\boldsymbol{y},\boldsymbol{X}]|\boldsymbol{s},\boldsymbol{X})|\boldsymbol{X}\right\} + \text{Var}\left\{\text{E}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s},\boldsymbol{X})|\boldsymbol{X}\right\} \\
&= \frac{1}{n^2 r}\boldsymbol{M}_X^{-1}\left[\sum_{i=1}^{n}\left\{p_i(\hat{\boldsymbol{\alpha}}_{MLE}) - 2p_i(\hat{\boldsymbol{\alpha}}_{MLE})p_i(\hat{\boldsymbol{\beta}}_{MLE}) + p_i(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}\boldsymbol{x}_i\boldsymbol{x}_i^T\left(\frac{1}{\pi_i} - 1\right)\right]\boldsymbol{M}_X^{-1} \\
&\quad + \frac{1}{n^2}\boldsymbol{M}_X^{-1}\sum_{i=1}^{n} p_i(\hat{\boldsymbol{\alpha}}_{MLE})\left\{1 - p_i(\hat{\boldsymbol{\alpha}}_{MLE})\right\}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{M}_X^{-1},
\end{aligned}
\tag{3.2}
$$

where $\boldsymbol{\alpha}$ is the vector of coefficients from the logistic regression model of y_i regressed on $\{s_i, x_i\}$, $\text{logit}\left\{p(y_i = 1|s_i,\boldsymbol{x}_i)\right\} = (s_i,\boldsymbol{x}_i)^T\boldsymbol{\alpha}$.

Proceeding with minimization of $\text{trace}\left[\text{Var}\left\{(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})|\boldsymbol{X}\right\}\right]$, we arrive at the following result.

Proposition 1: If the subsampling probability in Algorithm 1 is set to

$$\pi_{i,s\text{AUG}} = \frac{\sqrt{\left\{p_i(\hat{\boldsymbol{\alpha}}_{MLE}) - 2p_i(\hat{\boldsymbol{\alpha}}_{MLE})p_i(\hat{\boldsymbol{\beta}}_{MLE}) + p_i(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}}||\boldsymbol{M}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_i||}{\sum_{j=1}^{n}\sqrt{\left\{p_j(\hat{\boldsymbol{\alpha}}_{MLE}) - 2p_j(\hat{\boldsymbol{\alpha}}_{MLE})p_j(\hat{\boldsymbol{\beta}}_{MLE}) + p_j(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}}||\boldsymbol{M}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_j||},$$

then the asymptotic approximation error of $\tilde{\boldsymbol{\beta}}$, defined as $\text{trace}\left[\text{Var}\left\{(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE})|\boldsymbol{X}\right\}\right]$, will attain its minimum (see Section S1.2 for proof).

Note that these weights include $\hat{\boldsymbol{\alpha}}_{MLE}$ and $\hat{\boldsymbol{\beta}}_{MLE}$, which require knowledge of the gold-standard outcome to estimate. We therefore propose the following two-step algorithm for implementation, which uses a preliminary sample to provide pilot estimates of $\hat{\boldsymbol{\alpha}}_{MLE}$ and $\hat{\boldsymbol{\beta}}_{MLE}$ for use in constructing the proposed weights, $\pi_{i,s\text{AUG}}$, as shown below.

Algorithm 2: Surrogate-augmented sampling for outcome validation

---

1. **Step One**: r_1 observations are selected as a relatively non-informative sample, with weights equal to $1/n$. For selected observations, chart review is performed to uncover the true outcome, y, and the sampled data points denoted as $\boldsymbol{O}_i^* = (\boldsymbol{x}_i^*, s_i^*, y_i^*, \pi_i^*)$, for $i = 1, \ldots, r_1$. Pilot parameter estimates for $\hat{\boldsymbol{\alpha}}_{MLE}$ and $\hat{\boldsymbol{\beta}}_{MLE}$ are calculated using the r_1 observations.
2. **Step Two**: Pilot parameter estimates from step 1 are plugged into the more informative weights, $\pi_{i,s\text{AUG}}$. These weights are used to sample r_2 individuals in the second step for further chart review. Estimation of $\boldsymbol{\beta}$ proceeds through weighted estimation using r_2, as outlined in Algorithm 1.
3. **Step Three**: Denote the pilot estimate for $\boldsymbol{\beta}$ from step 1 as $\tilde{\boldsymbol{\beta}}_1$, and the estimate from step 2 as $\tilde{\boldsymbol{\beta}}_2$. Use inverse variance weighting to combine the estimates as follows:

$$\tilde{\boldsymbol{\beta}} = \left(\tilde{\boldsymbol{V}}_1^{-1} + \tilde{\boldsymbol{V}}_2^{-1}\right)^{-1}\left(\tilde{\boldsymbol{V}}_1^{-1}\tilde{\boldsymbol{\beta}}_1 + \tilde{\boldsymbol{V}}_2^{-1}\tilde{\boldsymbol{\beta}}_2\right) \qquad \tilde{\boldsymbol{V}} = \left(\tilde{\boldsymbol{V}}_1^{-1} + \tilde{\boldsymbol{V}}_2^{-1}\right)^{-1}$$

$$\tilde{\boldsymbol{V}}_1 = \tilde{\boldsymbol{M}}_{x,1}^{-1}\left\{\frac{1}{r_1^2}\sum_{i=1}^{r_1}\left(y_i^* - p_i^*(\tilde{\boldsymbol{\beta}}_1)\right)^2\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}\right\}\tilde{\boldsymbol{M}}_{x,1}^{-1}$$

$$\tilde{\boldsymbol{V}}_2 = \tilde{\boldsymbol{M}}_{x,2}^{-1}\left\{\frac{1}{n^2r_2^2}\sum_{i=1}^{r_2}\frac{1}{\pi_i^{*2}}\left(y_i^* - p_i^*(\tilde{\boldsymbol{\beta}}_2)\right)^2\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}\right\}\tilde{\boldsymbol{M}}_{x,2}^{-1}$$

$$\tilde{\boldsymbol{M}}_{x,1} = \frac{1}{r_1}\sum_{i=1}^{r_1}p_i^*(\tilde{\boldsymbol{\beta}}_1)(1 - p_i^*(\tilde{\boldsymbol{\beta}}_1))\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}, \qquad \tilde{\boldsymbol{M}}_{x,2} = \frac{1}{nr_2}\sum_{i=1}^{r_2}\frac{1}{\pi_i^*}p_i^*(\tilde{\boldsymbol{\beta}}_2)(1 - p_i^*(\tilde{\boldsymbol{\beta}}_2))\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}$$

Remark 1: Convergence issues and biased model estimates may result when fitting the models for $p(y_i|s_i, \boldsymbol{x_i})$ and $p(y_i|\boldsymbol{x_i})$ in small step 1 samples with rare events, preventing successful construction of the weights. To overcome this, we recommend using the Firth adjustment to the weighted log-likelihood in small samples (Firth, 1993).

Remark 2: Previous work on optimal sampling for logistic regression performs weighted estimation on the combined step 1 and step 2 samples. However, it was assumed that the step 1 sample size is negligible relative to the step 2 sample size, such that $r_1 = o(r_2^{1/2})$ (Wang, Zhu, and Ma, 2018). However, given the total study budget in a measurement constrained setting this may not be realistic, particularly if the model is large and a large r_1 is required for preliminary estimation of the parameters. Data combination when $r_1 > o(r_2^{1/2})$ can lead to bias in the final estimates, as showing consistency $\tilde{\boldsymbol{\beta}}$ for $\hat{\boldsymbol{\beta}}_{MLE}$ requires that $\sum_{i=1}^{n} \pi_i = 1$. Since outcome validation is costly and it is desirable to make use of all chart reviewed data, we propose an inverse variance weighted estimator that combines the estimates of $\boldsymbol{\beta}$ from step 1 (denoted as $\tilde{\boldsymbol{\beta}}_1$) and step 2 samples (denoted as $\tilde{\boldsymbol{\beta}}_2$), as shown in Algorithm 2. We assume that the overlap between the step 1 and step 2 samples is negligible (reasonable if $(r_1, r_2 << n)$). We also assume that the covariance between $\tilde{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_2$ induced by the incorporation of $\tilde{\boldsymbol{\beta}}_1$ in the weights for step 2 (where estimation of $\tilde{\boldsymbol{\beta}}_2$ then proceeds via a pseudo-likelihood) is not substantial (see results from Parke (1986)).

Remark 3: Intuitively, if $\hat{\boldsymbol{\alpha}}_{MLE}$ can be estimated with precision in the first step of sampling, then we can expect $\pi_{i,s\mathrm{AUG}}$ to perform similarly to $\pi_{i,y\mathrm{OBS}}$. This is because each individual's residual term in the weights, represented by $\sqrt{\left\{p_i(\hat{\boldsymbol{\alpha}}_{MLE}) - 2p_i(\hat{\boldsymbol{\alpha}}_{MLE})_i(\hat{\boldsymbol{\beta}}_{MLE}) + p_i(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}}$, will approximate $\sqrt{E\left\{(y_i - p_i(\hat{\boldsymbol{\beta}}_{MLE}))^2|s_i, \boldsymbol{x}_i\right\}}$ as $r \to \infty$, which by Jensen's inequality is $\geq E\left\{\sqrt{(y_i - p_i(\hat{\boldsymbol{\beta}}_{MLE}))^2}|s_i, \boldsymbol{x}_i\right\} = E\left\{|y_i - p_i(\hat{\boldsymbol{\beta}}_{MLE})||s_i, \boldsymbol{x}_i\right\}$. Since this relationship holds pointwisely 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.

Surrogate substitution weights

As a simpler alternative, we propose using weights that are A-optimal for the logistic regression of the surrogate outcome, s_i, on $\boldsymbol{x}_i$,

$$\text{logit}\left\{p(s_i = 1|\boldsymbol{x}_i)\right\} = \boldsymbol{x}_i^T\boldsymbol{\gamma}. \tag{3.3}$$

To understand the motivation behind this, we seek to provide an understanding of the relationship between models (3.3) and (3.1). Under perfect sensitivity and specificity, $\hat{\boldsymbol{\gamma}}$ will approximate $\hat{\boldsymbol{\beta}}$. Furthermore, the higher the correlation between y_i and s_i, the more likely an influential observation for the estimation of $\hat{\boldsymbol{\gamma}}$ will also be influential for estimation of $\hat{\boldsymbol{\beta}}$. Indeed, we can show geometrically that a portion of the influence an observation has for estimating $\hat{\boldsymbol{\beta}}$ can be modeled using the influence it has for estimating $\hat{\boldsymbol{\gamma}}$, and this portion is quantifiable using a projection of the influence function for $\hat{\boldsymbol{\gamma}}$, $\varphi^*(\hat{\boldsymbol{\gamma}}_{MLE})$, onto the tangent space spanned by the influence function for $\hat{\boldsymbol{\beta}}$, $\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})$ (see Section S1.3). This results in the following proposition.

Proposition 2: The influence of observation i on estimating $\hat{\boldsymbol{\beta}}_{MLE}$ is partially explained by its influence on estimation of $\hat{\boldsymbol{\gamma}}_{MLE}$. This portion is represented by the following quantity.

$$\Pi\left\{\varphi^*(\hat{\boldsymbol{\gamma}}_{MLE})|\wedge_{\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})}\right\} = \boldsymbol{Q}_X^{-1}\boldsymbol{A}_X\boldsymbol{M}_X^{-1}\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n}, \tag{3.4}$$

where $\wedge_{\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})}$ denotes the tangent space spanned by the influence function for $\hat{\boldsymbol{\beta}}$,

$$\boldsymbol{Q}_X = \frac{1}{n}\sum_{i=1}^{n} p_i^*(\hat{\boldsymbol{\gamma}}_{MLE})\left\{1 - p_i^*(\hat{\boldsymbol{\gamma}}_{MLE})\right\}\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T},$$

$$\boldsymbol{A}_X = \frac{1}{n}\sum_{i=1}^{n}\left\{p_i(y_i s_i|\boldsymbol{x}_i) - p_i^*(\hat{\boldsymbol{\gamma}}_{MLE})p_i^*(\hat{\boldsymbol{\beta}}_{MLE})\right\}\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}.$$

Proposition 2 offers a geometric representation of the component of $\varphi^*(\hat{\boldsymbol{\gamma}}_{MLE})$ that runs along the direction of $\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})$. Note that this quantity effectively scales the influence function for $\hat{\boldsymbol{\beta}}_{MLE}$ by $\boldsymbol{Q}_X^{-1}\boldsymbol{A}_X$, where $\boldsymbol{A}_X$ is a measure of covariance between y_i and s_i, conditional on $\boldsymbol{x}_i$. If $\boldsymbol{Q}_X^{-1}\boldsymbol{A}_X$ is high, then an observation that is informative for estimating $\hat{\boldsymbol{\gamma}}_{MLE}$ will also be informative in estimating $\hat{\boldsymbol{\beta}}_{MLE}$. On the other hand, if y_i and s_i are uncorrelated (conditional on $\boldsymbol{x}_i$), then $\boldsymbol{A}_X$

reduces to 0, and $\hat{\gamma}_{MLE}$ bears no relationship to $\hat{\boldsymbol{\beta}}_{MLE}$.

Based on the intuition from Proposition 2, we propose the following weights, which are A-optimal for estimation of $\tilde{\gamma}$ (based on Wang, Zhu, and Ma (2018)'s approach; see Section S1.4), for use in Algorithm 1,

$$\pi_{i,s\text{SUB}} = \frac{|s_i - p_i(\hat{\gamma}_{MLE})||\boldsymbol{Q}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_i||}{\sum_{j=1}^{n}|s_j - p_j(\hat{\gamma}_{MLE})||\boldsymbol{Q}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_j||},$$

where $\boldsymbol{Q}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} p_i(\hat{\gamma}_{MLE})(1-p_i(\hat{\gamma}_{MLE}))\boldsymbol{x}_i\boldsymbol{x}_i^T$. According to Proposition 2, if $\boldsymbol{Q}_X^{-1}\boldsymbol{A}_X$ is large, $\pi_{i,s\text{SUB}}$ may be superior to simple uniform sampling.

Since s_i is observed for all individuals, $\hat{\gamma}_{MLE}$ can be estimated using the full sample. Therefore, sampling with $\pi_{i,SUB}$ can be implemented in a single step, as shown in Algorithm 3. This is unlike $\pi_{i,AUG}$, which requires validating a preliminary sample in order to construct the weights. This holds two clear advantages. First, $\hat{\gamma}_{MLE}$ and therefore $\pi_{i,s\text{SUB}}$ can be estimated with high precision. Second, it enables all observations that are chart reviewed to be selected using informative weights, as opposed to a portion being selected with uniform sampling. However, if s_i is poorly correlated with y_i, then $\pi_{i,s\text{SUB}}$ will be very far from the optimal weights, $\pi_{i,y\text{OBS}}$, approaching non-informativeness (equivalent to uniform sampling) as $\text{corr}(\boldsymbol{s},\boldsymbol{y}) \to 0$. We explore this bias-variance tradeoff between $\pi_{i,s\text{AUG}}$ and $\pi_{i,s\text{SUB}}$ in simulations.

Algorithm 3: Surrogate-substitution sampling for outcome validation

---

1. $\hat{\gamma}_{MLE}$ is calculated using all n observations, and is used to construct the weights $\pi_{i,s\text{SUB}}$. These weights are then used to sample r individuals. For selected observations, chart review is performed to uncover the true outcome, y, and the sampled data points denoted as $\boldsymbol{O}_i^* = (\boldsymbol{x}_i^*, s_i^*, y_i^*, \pi_i^*)$, for $i = 1,\ldots,r$. Estimation of $\tilde{\boldsymbol{\beta}}$ proceeds through weighted estimation using the r observations as in Algorithm 1.

3.3. Simulation study

We evaluate how well each set of surrogate-assisted weights approximates the optimal weights $\pi_{i,y\text{OBS}}$ (for the ideal setting with y_i observed), as well as their performance in estimation of $\boldsymbol{\beta}$, a vector of 8 coefficients including the intercept, with true values equal to $(0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$, in the logistic regression model of y_i on $\boldsymbol{x}_i$. We consider and compare model estimates using the

optimal weights $\pi_{i,y\text{OBS}}$, our proposed weights, π_{isAUG} and $\pi_{i,s\text{SUB}}$, weights from Zhang, Ning, and Ruppert (2019) denoted as $\pi_{i,y\text{MISS}}$, simple uniform sampling (least informative), and the full data MLE. Three of the weights, $\pi_{i,y\text{OBS}}$, $\pi_{i,s\text{SUB}}$ and uniform weights, are implemented as single step sampling schemes, since the weights can be constructed using all of the observed data, while $\pi_{i,y\text{MISS}}$ and $\pi_{i,s\text{AUG}}$ require a two-step approach in which a preliminary sample is validated to obtain pilot estimates of certain parameters.

We consider multiple covariate distributions as described in Wang, Zhu, and Ma (2018) and Zhang, Ning, and Ruppert (2019), including:

- **zeroMean**. $\boldsymbol{x}$ follows a multivariate normal distribution with constant variance, defined as $MVN(\mathbf{0}, \boldsymbol{\Sigma})$, 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. .

- **nonzeroMean**. $\boldsymbol{x}$ follows a multivariate normal distribution centered away from 0, which induces moderate imbalance in the outcome such that $\sim 21\%$ of observations are cases. $\boldsymbol{x} \sim MVN(-\mathbf{0.8}, \boldsymbol{\Sigma})$, $\Sigma_{ij} = 0.5^{i \neq j}$.

- **unequalVar**. $\boldsymbol{x}$ follows a multivariate normal distribution with unequal variances. i.e. $\boldsymbol{x} \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\Sigma}_{ii} = 1/i^2$ for $i = 1, ..., 7$, and the off-diagonal entries equal to $\Sigma^*_{ij} = 0.5$ for $i \neq j$.

- **rareEvent**. Like `nonzeroNormal`, `rareEvent` has a covariate distribution centered away from 0, but one that induces more extreme imbalance in the outcomes ($\sim 5\%$ cases). $\boldsymbol{x} \sim MVN(-\mathbf{1.6}, \boldsymbol{\Sigma})$, $\Sigma_{ij} = 0.5^{i \neq j}$.

- **mixNormal**. $\boldsymbol{x}$ follows a bimodal distribution that is the mixture of two multivariate normal distributions ($0.5N(\mathbf{1}, \boldsymbol{\Sigma})$ and $0.5N(-\mathbf{1}, \boldsymbol{\Sigma})$), $\Sigma_{ij} = 0.5^{i \neq j}$.

- **Exp**. Each component of $\boldsymbol{x}$ follows an exponential distribution with a rate parameter of 2. The covariates are uncorrelated in this setting.

In performing a concordance analysis between the proposed surrogate-assisted weights ($\pi_{i,s\text{AUG}}$ and $\pi_{i,s\text{SUB}}$) and the true optimal weights, we perform a single simulation under each of the covariate distributions described above, with total dataset size $n = 10,000$ and step 1 sample size ranging

from $r_1 = 100, ..., 1000$ for $\pi_{i,s\text{AUG}}$. We calculate the mean-squared error for each of the estimated weights in comparison to $\pi_{i,y\text{OBS}}$, defined as,

$$MSE_{s\text{AUG}} = \frac{1}{n}\sum_{i=1}^{n}(\pi_{i,s\text{AUG}} - \pi_{i,y\text{OBS}})^2$$

and

$$MSE_{s\text{SUB}} = \frac{1}{n}\sum_{i=1}^{n}(\pi_{i,s\text{SUB}} - \pi_{i,y\text{OBS}})^2$$

For studying the performance of the proposed sampling designs on model efficiency, we perform $S = 500$ replications under each covariate distribution. For each replicate, a dataset of $n = 10,000$ is generated, and the total subsample size to be validated ranges from $r = 800, ..., 1600$. For the two-step approaches only, r includes a uniformly selected step 1 sample of size $r_1 = \{200, 600\}$ 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,

$$MSE_{\boldsymbol{\beta}} = \frac{1}{S}\sum_{s=1}^{S}||\tilde{\boldsymbol{\beta}}_s - \boldsymbol{\beta}||^2,$$

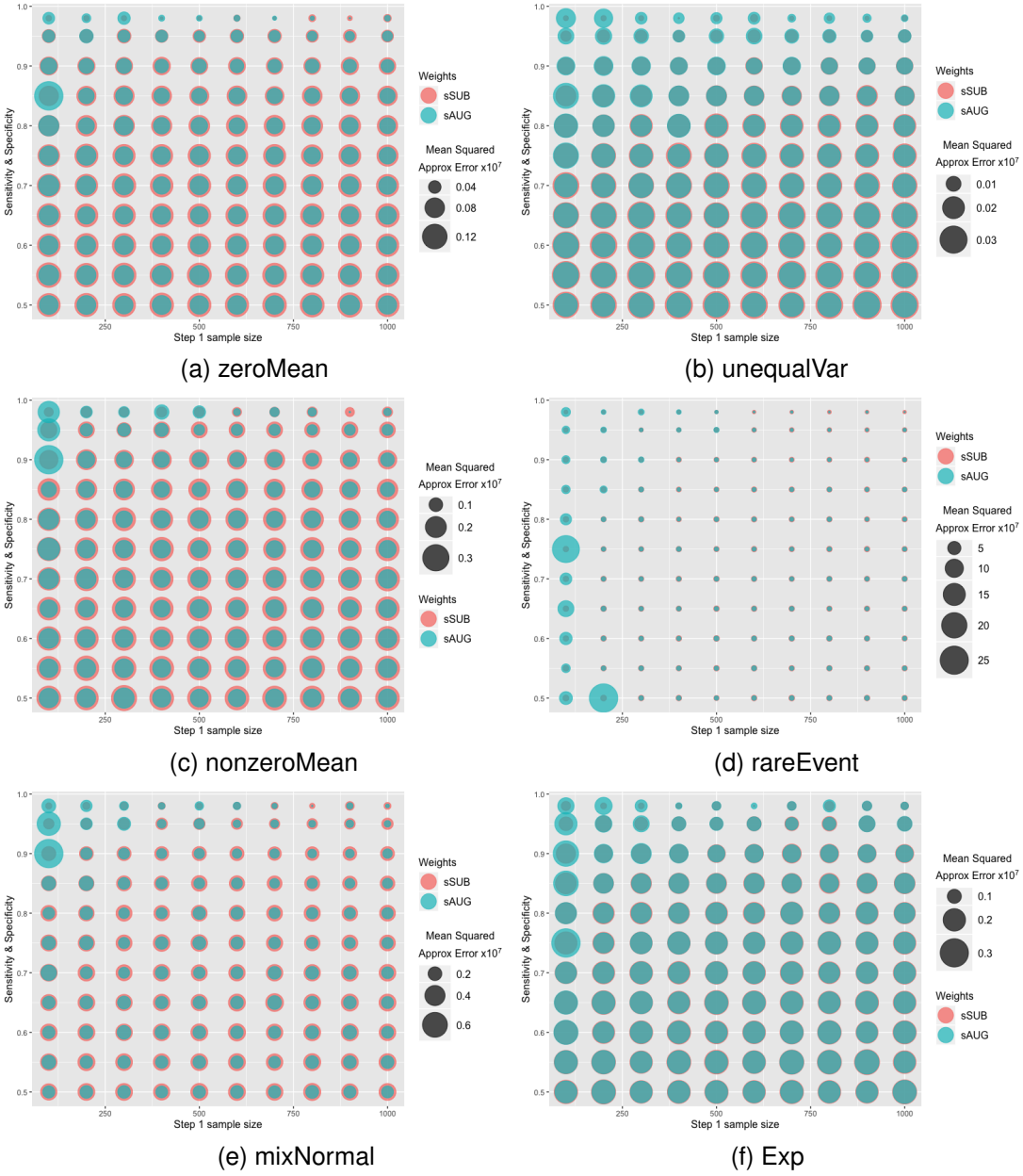
where $\boldsymbol{\beta}$ is the true parameter vector, and $\tilde{\boldsymbol{\beta}}_s$ is the estimate from the s^{th} replicate, and S is the number of replications.

Note that for the `rareEvent` setting, the Firth adjustment (Firth, 1993) is used when fitting logistic regression models to step 1 samples of size $r_1 = 200$ or smaller, to achieve high convergence rates for constructing the $\pi_{i,s\text{AUG}}$ weights.

3.3.1. Concordance analysis

Figure 3.1 plots the approximation errors of the proposed weights, $\pi_{s\text{AUG}}$ (blue) and $\pi_{i,s\text{SUB}}$ (orange), when each are compared to the true optimal weights, $\pi_{i,y\text{OBS}}$. The plots feature a range values for the step 1 sample size (r_1) for $\pi_{i,s\text{AUG}}$ (between 100 and 1000) and as well as sensitivity and specificity (which are set to be equal in all settings). Recall that $\pi_{i,s\text{SUB}}$ is unaffected by r_1, as the full dataset is used to construct the weights. Size of the bubble corresponds to the magnitude of the mean-squared approximation error compared to $\pi_{i,y\text{OBS}}$. Each of panels (a) through (f) displays results for one of the covariate distribution settings described above.

Figure 3.1: Bubble plot of approximation errors for $\pi_{i,s\text{AUG}}$ and $\pi_{i,s\text{SUB}}$ compared to $\pi_{i,y\text{OBS}}$, under various covariate distributions, sensitivity & specificity (set to be equal), and step 1 sample size (for $\pi_{i,s\text{AUG}}$) with $n = 10,000$. Approximation errors are small since $\sum_{i=1}^{n} \pi_i = 1$; they are multiplied by 10^7 for display purposes



(a) zeroMean (b) unequalVar

(c) nonzeroMean (d) rareEvent

(e) mixNormal (f) Exp

We see in all panels that $\pi_{i,s\text{AUG}}$ offers similar or better empirical approximation to the true optimal weights ($\pi_{i,y\text{OBS}}$) than $\pi_{i,s\text{SUB}}$ when sensitivity and specificity are low to moderate, and also when the step 1 sample (r_1) used to estimate $\hat{\boldsymbol{\beta}}_{MLE}$ and $\hat{\boldsymbol{\alpha}}_{MLE}$ is large. However, in very small step 1 samples (where r_1 may not be large enough to obtain precise pilot estimates) and under high

sensitivity and specificity (which can lead to collinearity issues in fitting logit$\{p(y_i|s_i, \boldsymbol{x_i})\}$), the augmented weights have poor approximation to the optimal weights compared to $\pi_{i,s\text{SUB}}$. Substitution weights may also be beneficial when the covariates have different variances (panel (c)) or follow and exponential distribution (panel (f)).

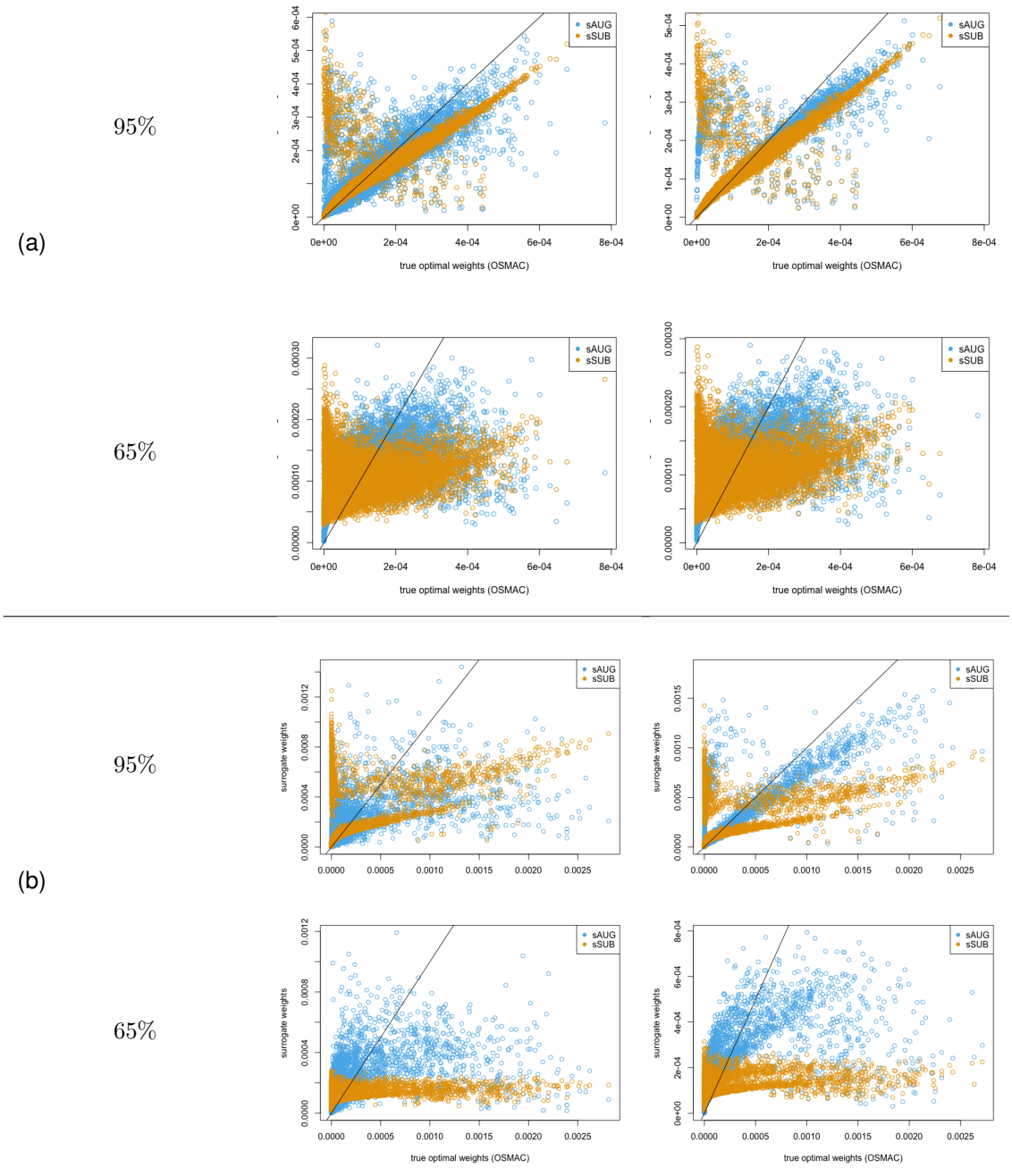
We provide further granularity in our analysis of the weights. Using a single set of simulated data under the `zeroMean` and `rareEvent` covariate distribution settings, step 1 sample size $r_1 = 200, 5000$ (for $\pi_{i,s\text{AUG}}$), sensitivity $se = \{0.95, 0.65\}$, and specificity $sp = \{0.95, 0.65\}$, we plot the individual estimated weights against their corresponding $\pi_{i,y\text{OBS}}$. Under perfect concordance between the weights, we would expect the points to fall along a diagonal line. A handful of these settings are shown in Figure 3.2. 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 $\pi_{i,s\text{AUG}}$ offers closer approximation to $\pi_{i,y\text{OBS}}$ on average, as points are scattered more evenly about the line $y = x$. However, with a small step 1 sample, the $\pi_{i,s\text{AUG}}$ weights are estimated with lower precision, and so have greater variability compared to $\pi_{i,s\text{SUB}}$ (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 3.2 that if sensitivity and specificity are low (eg., 65%), $\pi_{i,s\text{AUG}}$ 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 $\pi_{i,y\text{OBS}}$, as there is stronger concordance among the larger weights. On the other hand, the substitution weights, $\pi_{i,s\text{SUB}}$ follow a near random distribution when sensitivity and specificity are low, particularly if the event is rare.

3.3.2. Efficiency analysis

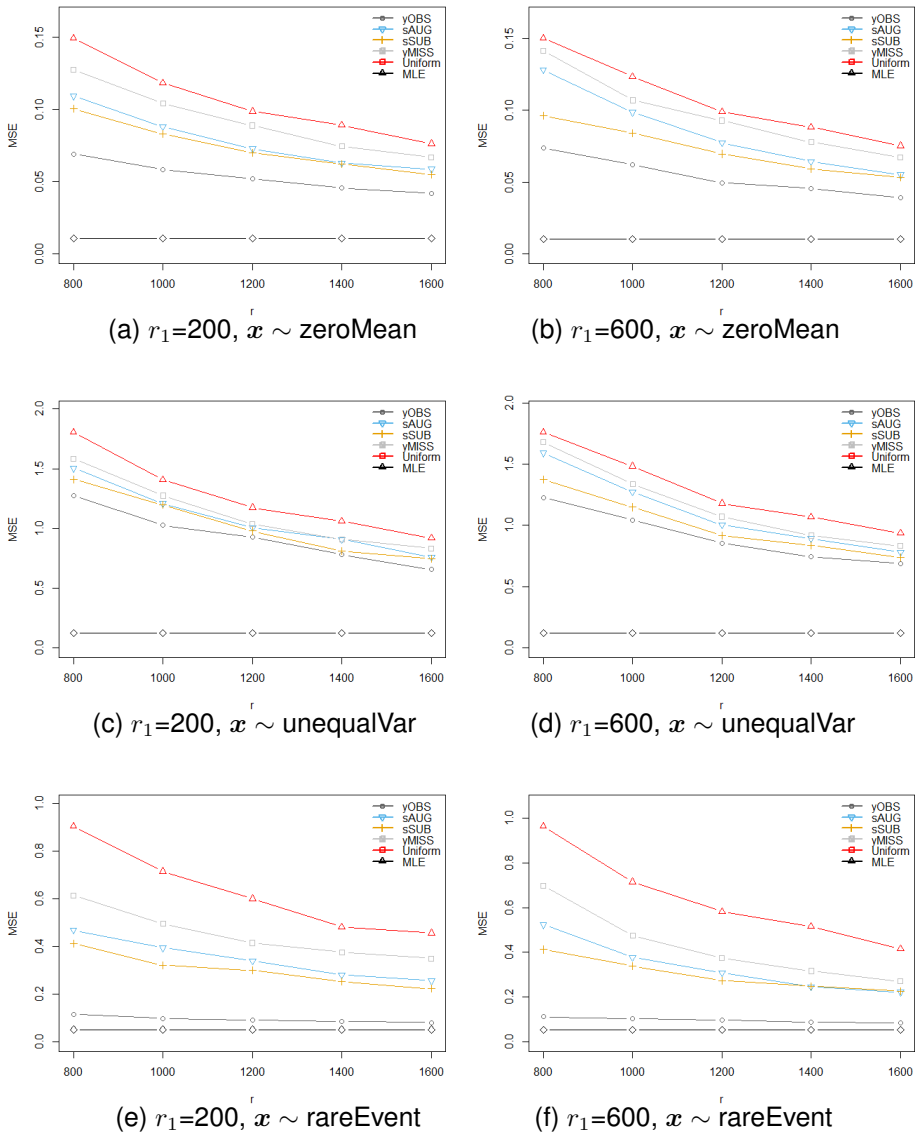
Figures 3.3 and 3.4 show the mean squared approximation error (MSE) results for $\tilde{\boldsymbol{\beta}}$ using various weights, including $\pi_{i,y\text{OBS}}$, $\pi_{i,s\text{AUG}}$, $\pi_{i,s\text{SUB}}$, weights from Zhang, Ning, and Ruppert (2019) denoted as $\pi_{i,y\text{MISS}}$, simple uniform sampling $(1/n)$, and the full data MLE. We consider settings of differential misclassification with moderately high sensitivity and specificity ($(se, sp)_{x_1 \leq 0.4} = (0.95, 0.90)$, $(se, sp)_{x_1 > 0.4} = (0.85, 0.80)$) (see Figure 3.3), and low sensitivity and specificity ($(se, sp)_{x_1 \leq 0.4} = (0.70, 0.65)$, $(se, sp)_{x_1 > 0.4} = (0.60, 0.55)$) (see Figure 3.4), as well as varying the step 1 sample size ($r_1 = 200, 600$) when a two-step sampling scheme is used ($\pi_{i,s\text{AUG}}$ and $\pi_{i,y\text{MISS}}$ only). Results

Figure 3.2: Concordance plots of $\pi_{i,s\text{AUG}}$ and $\pi_{i,s\text{SUB}}$, compared to $\pi_{i,y\text{OBS}}$, under (a) **zeroMean** covariate distribution and (b) **rareEvent** distribution, with $r_1 = 200, 5000$ (for $\pi_{i,s\text{AUG}}$) and sensitivity and specificity both equal to 95% or 65%. $n = 10,000$



are displayed for three of the covariate settings described (zeroMean, unequalVar, rareEvent), with results for the remaining settings included in Figures S1 and S2 in the supplementary material.

Figure 3.3: Empirical mean squared error of $\tilde{\beta}$ using different sampling weights, over 500 replicates. $(se, sp)_{x_1 \leq 0.4} = (0.95, 0.90)$, $(se, sp)_{x_1 > 0.4} = (0.85, 0.80)$, $n = 10,000$



(a) r_1=200, $\boldsymbol{x} \sim$ zeroMean (b) r_1=600, $\boldsymbol{x} \sim$ zeroMean

(c) r_1=200, $\boldsymbol{x} \sim$ unequalVar (d) r_1=600, $\boldsymbol{x} \sim$ unequalVar

(e) r_1=200, $\boldsymbol{x} \sim$ rareEvent (f) r_1=600, $\boldsymbol{x} \sim$ rareEvent

As expected, the weights by Wang, Zhu, and Ma (2018), $\pi_{i,y\text{OBS}}$, achieved optimal efficiency in all settings. Uniform sampling was least efficient in most settings, and the two proposed weights ($\pi_{i,s\text{AUG}}$ and $\pi_{i,s\text{SUB}}$) achieved MSEs in between the optimal weights for the setting where y_i is observed ($\pi_{i,y\text{OBS}}$) and the optimal weights for the setting where y_i is missing ($\pi_{i,y\text{MISS}}$). When sensitivity and specificity are high, the single-step substitution weighting scheme achieves a similar or often better (lower) MSE compared to the two-step augmented weights (see Figure 3.3). This

holds true with both small and large step 1 samples. Even though the larger $r_1 = 600$ leads to better construction of the $\pi_{i,s\mathrm{AUG}}$ weights, as observed in the concordance analysis, the 600 individuals are being uniformly (i.e. non-informatively) selected, whereas they can all be selected using more informative weights in single-step approaches such as surrogate substitution weighting, hence the better performance of $\pi_{i,s\mathrm{SUB}}$ compared to $\pi_{i,s\mathrm{AUG}}$ in the right-hand panel of Figure 3.3. Importantly, for rare event data, the subsample size needs to be twice as large under uniform sampling to achieve a similar MSE as when either of the surrogate-assisted sampling weights are used.

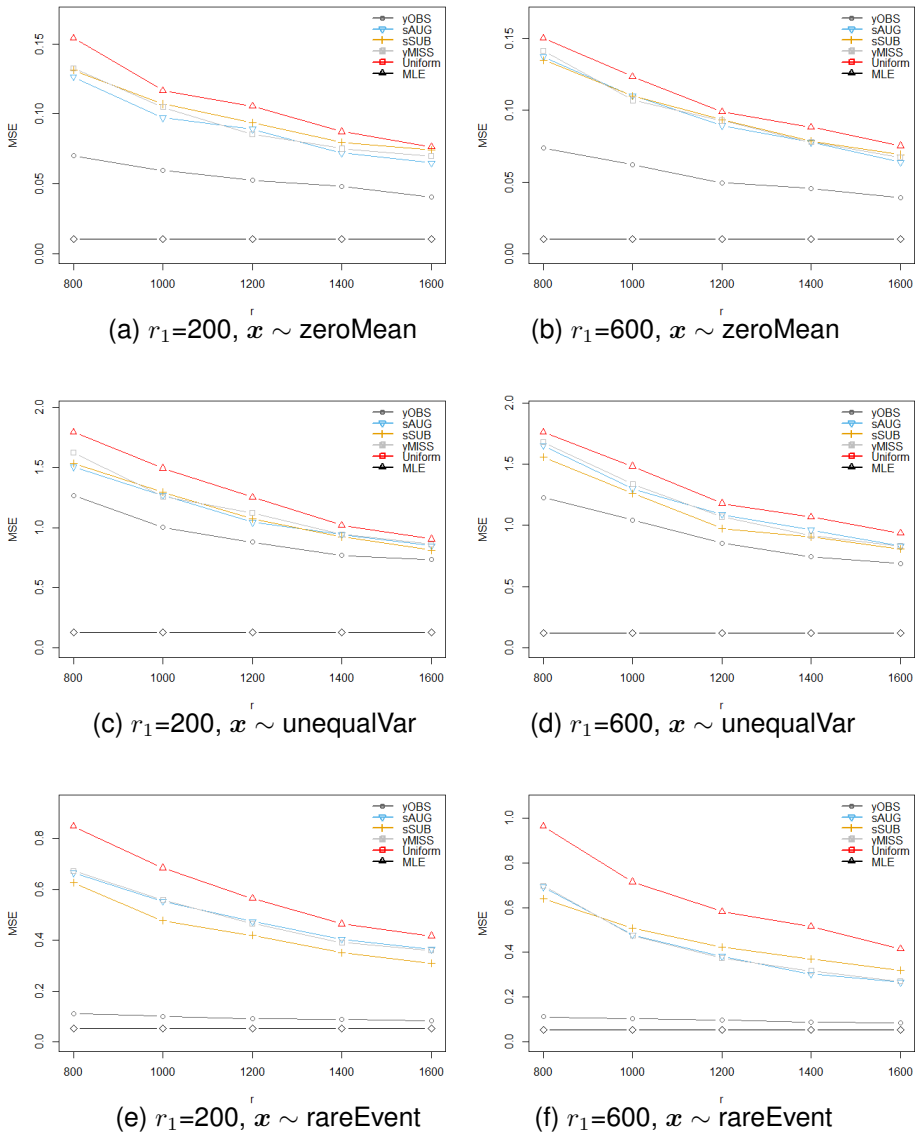
When sensitivity and specificity are low, the efficiency gains using the surrogate-assisted weights relative to uniform sampling are reduced, but remain substantial (for rare event data, uniform sampling still requires 1.3 times the subsample size to achieve similar efficiency). The performance of the two surrogate-assisted weights are also more similar, with $\pi_{i,s\mathrm{SUB}}$ slightly underperforming relative to $\pi_{i,s\mathrm{AUG}}$ for certain covariate distributions (see panels (a), (b) and (f) in Figure 3.4), and Figure S2 in the supplementary material. Also note that in the rare-event setting, the relative advantage of using $\pi_{i,s\mathrm{AUG}}$ (with large r_1) increases as the total subsample size increases. Furthermore, we observe that $\pi_{i,s\mathrm{AUG}}$ always results in a similar or slightly lower MSE than $\pi_{i,y\mathrm{MISS}}$. This is expected since $\pi_{i,y\mathrm{MISS}}$ and $\pi_{i,s\mathrm{AUG}}$ were 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 $\pi_{i,s\mathrm{AUG}}$ will reduce to $\pi_{i,y\mathrm{MISS}}$. Thus under low sensitivity and specificity, the surrogate augmented weights will offer similar or slightly more information to the weights that utilize $\boldsymbol{x}_i$ only, while the surrogate substitution weights may underperform relative to the weights that use $\boldsymbol{x}_i$ only.

Therefore, we propose the use of single-step surrogate substitution weights for settings where sensitivity and specificity of the surrogate outcome are expected to be relatively high, and also for rare event settings. However, the surrogate augmented weights are more robust to sensitivity and specificity, never underperforming compared to $\pi_{i,y\mathrm{MISS}}$, which are tailored for the worst-case scenario where only $\boldsymbol{x}_i$ is observed.

3.4. Application to BRAVA Study

Here we apply the candidate weights to an EHR dataset from the BRAVA study conducted at Kaiser Permanente Washington (KPWA) (Boudreau et al., 2014), which studied risk factors for second breast cancer events (SBCE) in women with a personal history of breast cancer. The

Figure 3.4: Empirical mean squared error of $\tilde{\beta}$ using different sampling weights, over 500 replicates. $(se, sp)_{x_1 \leq 0.4} = (0.70, 0.65)$, $(se, sp)_{x_1 > 0.4} = (0.60, 0.55)$, $n = 10,000$



(a) r_1=200, $\boldsymbol{x} \sim$ zeroMean (b) r_1=600, $\boldsymbol{x} \sim$ zeroMean

(c) r_1=200, $\boldsymbol{x} \sim$ unequalVar (d) r_1=600, $\boldsymbol{x} \sim$ unequalVar

(e) r_1=200, $\boldsymbol{x} \sim$ rareEvent (f) r_1=600, $\boldsymbol{x} \sim$ rareEvent

data consist of 3,152 women diagnosed with primary stage I-IIIB invasive breast cancer between 1993 and 2006. This dataset 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. (2012). In practice we expect only a subset of patients would have their outcomes clinically validated. To illustrate the use of our methods in the logistic regression

setting, we focus our analysis on the risk of an SBCE within 3 years of the index breast cancer diagnosis, excluding women who were lost to follow-up under the assumption of non-informative missingness. Our final dataset includes 2,813 women.

We seek to fit a model for the risk of an SBCE within 3 years of the index breast cancer diagnosis using the following predictors: patient's age at primary breast cancer diagnosis, SEER stage of the index cancer, and a categorical variable representing the combined estrogen (ER) and progesterone (PR) receptor status of the index cancer. We vary the total subsample size from $r = 600$ to 1000, and the step 1 sample size (for $\pi_{i,s\text{AUG}}$ and $\pi_{i,y\text{MISS}}$) across $r_1 = 200$ and 400. For each sample size setting, we draw $S = 500$ samples from the dataset, and calculate empirical mean squared approximation error (MSE) of $\tilde{\boldsymbol{\beta}}$ compared to the full data MLE, $\hat{\boldsymbol{\beta}}_{MLE}$. We further report the mean bias and model-based variance for the effect of the index breast cancer being ER/PR-negative, as compared to being ER-positive, on SBCE. Hormone therapy can be used to block ER-positive and/or PR-positive tumors, thereby slowing tumor growth and reducing the risk of recurrent cancer. Patients with ER/PR-negative tumors cannot benefit from this treatment. The full data MLE using the gold-standard outcome results in an estimate of $\beta_{ER/PR-} = 1.17$ (95% CI=0.84,1.51), or odds ratio $OR_{ER/PR-} = 3.23$ (95% CI=2.31,4.54), adjusting for age at index cancer diagnosis and SEER stage of index cancer. To perform a cost-analysis, we consider the scenario where the per-individual cost of measuring the gold-standard outcome is $50, and report the total number of unique records that are validated using each of the sampling schemes. As records are sampled with replacement, we expect that more informative weights would results in fewer records requiring chart review.

A total of 182 gold-standard SBCE events were observed (6.5%), while the phenotyping algorithm identified 166 cases. The computable phenotype has near perfect specificity (99%) and moderately high sensitivity (81%), though this may differ within patient subgroups. The results of our analysis are shown in Table 3.1. We see that uniform sampling with severely imbalanced outcome data resulted in larger variance in estimating $\beta_{ER/PR}$ compared to all informative sampling approaches, as well as a high empirical MSE of the model. Similar to the simulation study results, both surrogate-assisted weights achieved a lower MSE than $\pi_{i,y\text{MISS}}$, which only uses covariate information. In almost all settings studied, $\pi_{i,s\text{SUB}}$ resulted in a smaller model MSE and variance of $\beta_{ER/PR-}$ in comparison to $\pi_{i,s\text{AUG}}$, as well as a smaller number of total records validated. Only when $r_1 = 200$

Table 3.1: Empirical results for $\tilde{\boldsymbol{\beta}}$ over 500 replicates using BRAVA dataset ($n = 2813$)

		$r_1 = 200$			$r_1 = 400$		
		r			r		
Measure	Weights	600	1000	1400	600	1000	1400
Empirical MSE of $\tilde{\boldsymbol{\beta}}$ compared to $\hat{\boldsymbol{\beta}}_{MLE}$	$\pi_{i,s\text{SUB}}$	0.298	0.188	0.147	0.288	0.183	0.123
	$\pi_{i,s\text{AUG}}$	0.372	0.207	0.144	0.496	0.262	0.169
	$\pi_{i,y\text{MISS}}$	0.594	0.338	0.256	0.719	0.393	0.308
	$\pi_{i,uniform}$	23.116	6.870	0.844	15.562	4.986	1.343
Bias of $\tilde{\beta}_{ER/PR}$ compared to $\hat{\beta}_{ER/PR,MLE}$	$\pi_{i,s\text{SUB}}$	0.004	0.015	0.012	0.013	0.010	0.005
	$\pi_{i,s\text{AUG}}$	0.013	-0.004	-0.015	0.007	-0.003	-0.016
	$\pi_{i,y\text{MISS}}$	-0.019	-0.010	-0.013	0.024	0.001	-0.014
	$\pi_{i,uniform}$	-0.025	-0.003	-0.016	-0.027	-0.003	-0.019
Variance $\tilde{\beta}_{ER/PR}$	$\pi_{i,s\text{SUB}}$	0.064	0.040	0.030	0.064	0.040	0.029
	$\pi_{i,s\text{AUG}}$	0.075	0.045	0.032	0.090	0.049	0.033
	$\pi_{i,y\text{MISS}}$	0.118	0.073	0.053	0.122	0.075	0.054
	$\pi_{i,uniform}$	0.144	0.084	0.060	0.144	0.085	0.059
Number of validated records	$\pi_{i,s\text{SUB}}$	398	571	720	397	572	718
	$\pi_{i,s\text{AUG}}$	508	746	938	531	782	980
	$\pi_{i,y\text{MISS}}$	531	811	1045	538	824	1060
	$\pi_{i,uniform}$	541	841	1103	541	842	1102
Total cost ($50 per record)	$\pi_{i,s\text{SUB}}$	$19900	$28550	$36000	$19850	$28600	$35900
	$\pi_{i,s\text{AUG}}$	$25400	$37300	$46900	$26550	$39100	$49000
	$\pi_{i,y\text{MISS}}$	$26550	$40550	$52250	$26900	$41200	$53000
	$\pi_{i,uniform}$	$27050	$42050	$55150	$27050	$42100	$55100

and r is large did $\pi_{i,s\text{AUG}}$ achieve a similar MSE to $\pi_{i,s\text{SUB}}$, and still at substantially greater cost ($+\$10,900$) with 30% more records being validated. This is again due to the $\pi_{i,s\text{SUB}}$ weights not requiring a step 1 sample, thus allowing all r observations to be selected informatively, as opposed to 200 or 400 of the total budgeted records being selected with non-informative, uniform weights. The superior performance of $\pi_{i,s\text{SUB}}$ is also aided by the high accuracy of the phenotype, such that the weights $\pi_{i,s\text{SUB}}$ will closely approximate $\pi_{i,y\text{OBS}}$, as well as the low incidence of the outcome of interest.

It is important to note that the efficiency gains from all four informative sampling approaches were achieved using fewer validated observations compared to uniform sampling (due to sampling with replacement; if sampling occurred without replacement then the number of observations would be equal), illustrating the cost-effectiveness of such approaches. The least informative of the weights considered, $\pi_{i,y\text{MISS}}$, saves $150 to $2,900 in study costs for a given subsample size r when compared to uniform sampling, while $\pi_{i,s\text{SUB}}$ samples the fewest number of unique records given r, saving between $7,150 and $19,200 compared to uniform sampling. Notably, in order to achieve

a similar variance of $\tilde{\beta}_{ER/PR-}$ (~ 0.06) as when using the $\pi_{i,s\mathrm{SUB}}$ weights, uniform sampling needs to validate 2.7 times as many records, costing over $35,000 more than the surrogate sampling scheme.

We perform a sensitivity analysis to evaluate the performance of the weights under a lower accuracy of the surrogate phenotype. We modified the surrogate outcome to have 62% sensitivity and 68% specificity. Results are shown in Table S1 in the supplementary material. We observe that the variances and MSEs using both surrogate assisted weighting schemes have substantially increased when the sensitivity and specificity are lower, approaching the MSEs obtained using $\pi_{i,yMISS}$. We also see that $\pi_{i,sAUG}$ and $\pi_{i,sSUB}$ result in similar MSEs, with $\pi_{i,sAUG}$ yielding slightly lower variance estimates for $\tilde{\beta}_{ER/PR-}$ compared to $\pi_{i,sSUB}$. Importantly, both surrogate-assisted weights appear to be much less cost-effective under lower sensitivity and specificity of the surrogate outcome, with cost savings being reduced to a range of $1,050-$4,150 for $\pi_{i,sSUB}$ and $150-$3700 for $\pi_{i,sAUG}$. Thus the surrogate-assisted weights are most beneficial under high sensitivity and specificity of the phenotype, though minimal cost savings remain under a lower level of accuracy.

3.5. Discussion

Differential misclassification of EHR-derived phenotypes can lead to biased inference in EHR-based 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. In this paper we proposed a means of improving statistical efficiency in the estimated model parameters from the validation set, through the use of informative sampling weights.

We introduced two candidate sampling schemes for guiding chart review to obtain gold-standard EHR outcomes in measurement constrained settings using the surrogate, EHR-derived phenotype. These weights may also be used in settings where an auxiliary variable for the outcome is available, which would not be included as an outcome or predictor in the association model of interest. The first set of weights, called surrogate augmented weights, require a two-step framework, in which a smaller sample is first selected for outcome ascertainment in order to construct the weights. These are fairly robust to the level of accuracy of the surrogate outcome, but may not perform optimally if the step 1 size is very small or the true event is rare. The second set of weights,

surrogate substitution weights, do not require the gold standard outcome and can be computed with precision using the entire dataset. However, the substitution weights may offer little advantage over uniform sampling if the computable phenotype is weakly correlated with the gold standard outcome. We offered an illustration of this robutsness vs. efficiency tradeoff through simulation studies and application to the study of risk factors for secondary breast cancer events (SBCEs) in a KPWA dataset. We found that under fairly high sensitivity and specificity, the surrogate substitution weights offer the greatest efficiency gains, along with potentially substantial cost savings. However, under low sensitivity/specificity the efficiency gains are reduced, and the augmented weights, being more robust to phenotype accuracy, may be preferred over substitution weights.

It should be noted that phenotyping algorithms used in practice generally have moderate to high levels of accuracy, which favors the use of surrogate substitution weights for risk-association studies. However, the surrogate outcome need not be computationally derived through sophisticated algorithms. A simple proxy variable that is mildly informative of the outcome outcome (for example, ICD-10 code or a single biomarker) may still confer additional information for selecting a validation set compared to less informative weights that only use the covariates (developed by Zhang, Ning, and Ruppert (2019)). Surrogate augmented weights can be particularly useful in this setting.

These approaches are not without limitations. Notably, our inverse-variance weighted estimator for the two-step surrogate-augmented approach assumes that the estimates from the first-step and second-step samples are uncorrelated. While we expect their covariance to be small, particularly in large datasets, since their dependence is through the use of the step 1 sample estimate in the weights for step 2, it is not zero, and so our variance formulation proposed in Algorithm 2 may underestimate the true variance of $\tilde{\boldsymbol{\beta}}$. We encourage study of more robust formulations. Second, it is unknown how misspecification of the working models used to construct the weights may impact performance. We assumed a linear form when modeling $logit(p(s_i|\boldsymbol{x}_i))$ and $logit(p(y_i|s_i,\boldsymbol{x}_i))$, but this may be misspecified for certain datasets. The results in propositions 1-2 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 misspecified leading to suboptimality in the final validation sample estimates, though $\tilde{\boldsymbol{\beta}}$ should remain unbiased by virtue of inverse probability weighting. The robustness of these approaches to model misspecification is worthy of future investigation.

Another limitation is that sampling approaches that use leverage to measure the informativeness

of observations have the potential to oversample outliers in the data, which could have undue influence on the final estimates. However, as the weighted log-likelihood uses the inverse of the weights, such influence should be minimal. The use of inverse weighting for estimation is in itself a limitation of algorithmic leveraging, as observations with larger sampling weights end up contributing less to the weighted estimation procedure, thus reducing efficiency. Wang (2019) 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 *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. (2020), to further improve efficiency. An R package for implementing the proposed weights is forthcoming.

CHAPTER 4

AIM 3: SELECTION BIAS IN EHR-BASED STUDIES - ACHIEVING EXTERNAL VALIDITY IN SURROGATE-ASSISTED VALIDATION STUDY DESIGNS

4.1. Background

In Aim 2 we proposed surrogate-assisted sampling schemes to improve efficiency in estimation of logistic regression models in measurement constrained settings. While these study designs hold promise for more cost-effective outcome validation relative to non-informative, random sampling, their use has only been studied in settings where the study population is assumed to be representative of the target population. We wish to consider their use in more complex settings which may be encountered in EHR data. Specifically, we seek to address the external validity of estimates obtained using a validated sample of EHR data where selection bias may be present.

EHR data are not prospectively collected for research purposes, but rather they are obtained when patients interact with the healthcare system. As such, EHR data may be subject to selection bias if certain population subgroups are more likely than others to interact with healthcare systems, specifically EHR-supported institutions. For example, disparities exist in terms of access to technology and the 'digital divide' (Ibrahim, Charlson, and Neill, 2020) in a time where electronic and mobile health data are increasingly used in observational studies and for patient recruitment. Differences in uptake of EHR systems by clinics serving underserved groups (eg. rural, uninsured, those served by smaller providers) (Hamamura, Withy, and Hughes, 2017; Hing and Burt, 2009; Kemper, Uren, and Clark, 2006; Mack et al., 2016), the availability of patient-reported outcomes (Rathod and Wragg, 2018), and the quality of data on race/ethnicity (Lee, Grobe, and Tiro, 2016; López et al., 2017) also impact the representativeness of EHR data. Lack of representation is further compounded if studies link EHR data to external sources such as biobank data, where the lack of diversity in the data repositories is well documented. For example, 94.6% of participants in the UK Biobank are of white ethnic origins, with only 1.6% of participants of black or black-british ethnicity (Fry et al., 2017). This is compared to 80% and 3.4% of the UK population that are white and black respectively, based on the 2011 census. Furthermore, 68% of global enrollment in biobanks (which are increasingly linked to EHR data) consists of those with European ancestry (Abul-Husn

and Kenny, 2019). Incomplete records or those not included in large, integrated EHR systems may be excluded from observational studies or embedded pragmatic trials.

When EHR patient populations are unrepresentative of the target population of interest, inference made on an EHR sample cannot be generalized to the larger population. For example, if selection into the EHR sample is dependent on the outcome and/or a set of auxiliary variables that are also associated with the outcome, this can lead to biased model estimates relative to the true data generating mechanism in the population (Bower et al., 2017; Haneuse and Daniels, 2016). Furthermore, selection mechanisms that depend on an effect modifier (a variable over which the treatment effect of interest varies), can lead to different overall treatment effects in the study population compared to the target or source population. These forms of selection bias refer to a study's external validity. If the selection mechanism is known, inverse probability weighted estimation can correct for any bias due to selection. However, in practice selection probabilities must be estimated using external knowledge and/or data. Beesley and Mukherjee (2020) proposed a framework for jointly accounting for misclassification error and selection bias when gold-standard outcomes are unavailable. In the setting where study outcomes can be validated for a limited number of individuals, Zhang et al. (2020) proposed optimal sampling weights accounting for selection bias in a two-phase sampling framework (with the first phase corresponding to selection into the EHR database, and the second phase corresponding to selection into the study sample). However, this was geared towards estimation of the population mean (as opposed to a model-based approach) and assumed the outcome was unobserved in the first phase, rather than misclassified.

In this paper we jointly address outcome misclassificaton and selection bias in EHR-based association studies by combining efficient model-based sampling weights for outcome validation (developed by Zhang et al. (2020) and Marks-Anglin et al. (2021)) with methods to account for selection bias in the final subsample estimates. We focus on logistic regression for modelling risk-factor associations. When selecting a subsample from an EHR dataset on which to validate outcomes, the sampling weights can improve model precision relative to uniform sampling in the final association analysis. However, more adjustment is needed to account for selection bias in the EHR data relative to the target population (to improve external validity). Furthermore, we consider how selection bias impacts the ability of the sampling weights to improve statistical efficiency in the final model estimates, as the weights are model-based and constructed using pilot estimates derived from the

EHR study cohort. If the EHR cohort is a selective sample of the target population, the weights might be efficient for a biased model, rather than the true model of interest. We draw from the literature on calibration and inverse probability weighting for selection bias to improve our weighted estimation procedure and produce estimates that are generalizable to the target population. Our approach is then used to generalize the results of an association study on colon cancer recurrence performed using a validated subsample from Kaiser Permanente data. While Kaiser Permanente, a large managed care organization, operates in 8 states across the U.S., the data included in their study come from their Washington healthcare system only, where non-white racial-subgroups are underrepresented relative to the larger national population. Furthermore, a majority of Kaiser Permanente patients are employed and all are insured, thus excluding more vulnerable, uninsured populations from our analysis. We wish to transport the results of our analysis to the larger U.S. colon cancer patient population.

In Section 4.2.1, we introduce the two-phase sampling framwork for validation studies using EHR databases and define forms of selection bias. In Section 4.2.2, we outline existing sampling approaches for selecting informative subsamples for logistic regression, followed by inverse probability weighting methods to account for selection bias in Section 4.2.3. We perform a simulation study comparing the methods for sampling and adjustment for selection bias in Section 4.3. Finally, in Section 4.4 we illustrate the use of informative sampling and poststratification weights to perform an externally valid association study using a subsample of colon cancer recurrence data from Kaiser Permanente, as well as statistics from the Surveillance, Epidemiology, and End Results (SEER) Program at the National Cancer Institute.

4.2. Methods

4.2.1. Two-Phase Sampling Framework

To illustrate our setting we adapt a two-phase sampling framework shown in Zhang et al. (2020). Consider a target population of N individuals with outcome vector y, and an $N \times p$ covariate matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)^T$. We are interested in estimating the true disease model in the target population, represented by the following logistic regression model,

$$\text{logit}\{P(y_i = 1|\boldsymbol{x}_i)\} = \boldsymbol{x}_i\boldsymbol{\beta}. \tag{4.1}$$

Let $R_{i,1} \in \{0,1\}$ be a binary indicator representing inclusion in an EHR study cohort of size n. We assume that the probability of inclusion may depend on y and/or auxilliary variables $\boldsymbol{A} = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N)^T$ (i.e. $p(R_{i,1} = 1|y_i, \boldsymbol{a}_i)$) that are associated with $\boldsymbol{x}_i$, y_i or both. Some of these variables may be also included in the disease model, while others are not (we denote these as $\boldsymbol{A}^\wedge$). We assume that either individual level auxilliary data $\boldsymbol{a}_i$ or the distribution of $\boldsymbol{a}_i$ (denoted as $f(\boldsymbol{a}_i)$)) is available for the general population. We denote the selection or inclusion probability for an EHR cohort as follows,

$$P(R_{i,1} = 1|y_i, \boldsymbol{a}_i, \boldsymbol{\eta}) = \pi_1(y_i, \boldsymbol{a}_i, \boldsymbol{\eta}) \tag{4.2}$$

such that the disease model in the EHR study cohort has the following form:

$$\text{logit}\{P(y_i = 1|\boldsymbol{x}_i, R_{i,1} = 1)\} = \boldsymbol{x}_i\boldsymbol{\theta}. \tag{4.3}$$

We assume that $\pi_1(y_i, \boldsymbol{a}_i, \boldsymbol{\eta})$ is not known or controlled by the investigator in practice.

Furthermore, the outcome in the EHR cohort is potentially misclassified or missing, and instead we observe s_i (which can be a computable phenotype or a related proxy variable). We assume that we can perform a validation study for a small sample of size r ($r << n$), where we uncover the gold standard outcome y_i through chart review or some other validation procedure. Denote selection into the validation sample with indicator $R_{i,2} \in \{0,1\}$, where selection probabilities may be a function of $\boldsymbol{x}_i$ and/or s_i, as follows,

$$P(R_{i,2} = 1|\boldsymbol{x}_i, s_i, R_{i,1} = 1, \boldsymbol{\phi}) = \pi_2(\boldsymbol{x}_i, s_i, \boldsymbol{\phi}). \tag{4.4}$$

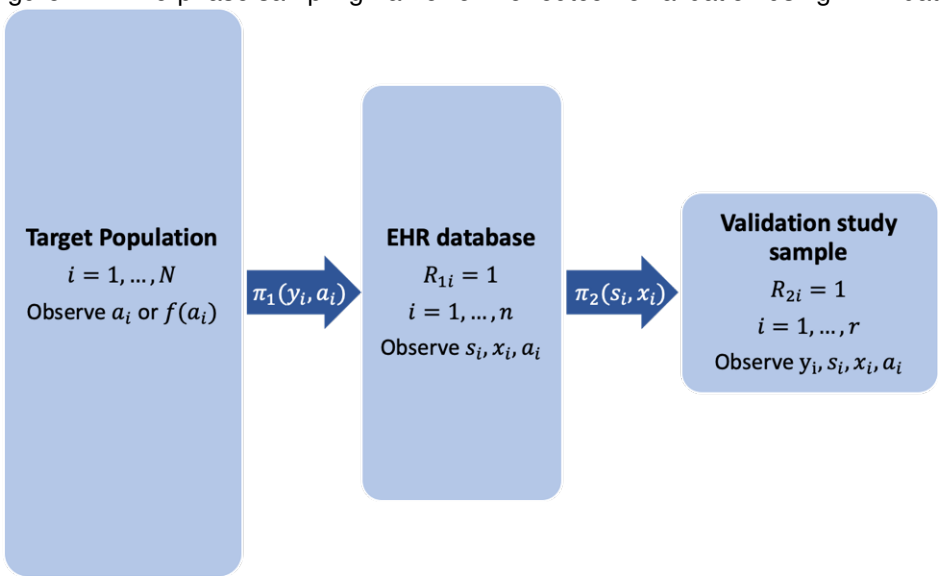
Here, ϕ includes pilot parameters estimates calculated using some or all of the EHR study cohort, as will be discussed in the next section. We assume these weights are computed and known to the investigator.

Finally, the disease model in the validation sample takes the following form:

$$\text{logit}\{P(y_i = 1|\boldsymbol{x}_i, R_{i,1} = 1, R_{i,2} = 1)\} = \boldsymbol{x}_i\boldsymbol{\theta}^*. \tag{4.5}$$

Thus sampling occurs in two-phases; first into the EHR dataset, and second into the validation sample. We denote the phase 1 sampling weights as $\pi_1(y_i, \boldsymbol{a}_i, \boldsymbol{\eta})$, and the phase 2 weights as

Figure 4.1: Two-phase sampling framework for outcome validation using EHR data



$\pi_2(s_i, \boldsymbol{x}_i, \phi)$. We illustrate this framework in Figure 4.2.1 below. In this setting, selection bias may be present such that $\boldsymbol{\theta}^* \neq \boldsymbol{\theta} \neq \boldsymbol{\beta}$, and a naïve analysis fit to the validation set is subject to selection bias if we are interested in inference on $\boldsymbol{\beta}$ (Beesley and Mukherjee, 2020).

Consider also that a_i could be an effect modifier of the relationship between $\boldsymbol{x}_i$ and y_i. For example, if a_i is binary, we may have that

$$\begin{aligned} \text{logit}\{P(y_i = 1 | \boldsymbol{x}_i, a_i = 0)\} &= \boldsymbol{x}_i \boldsymbol{\beta}_0, \\ \text{logit}\{P(y_i = 1 | \boldsymbol{x}_i, a_i = 1)\} &= \boldsymbol{x}_i \boldsymbol{\beta}_1. \end{aligned} \tag{4.6}$$

In this setting, the overall effect of $\boldsymbol{x}_i$ observed in a population depends on the distribution of a_i in that population. Therefore, if the EHR study cohort does not reflect the target population with respect to the distribution of a_i, then the estimated effect using the EHR cohort may be different from that in the target population (model 4.1).

Thus the first phase of sampling must be accounted for in order to provide valid inference in the second phase.

4.2.2. Validation sampling weights

Several design approaches have been proposed for the second phase of sampling in the framework shown in Figure 4.2.1. The simplest form for the phase 2 weights is $\pi_2 = 1/n$, also known as *uniform sampling*. However, these may not be the most efficient weights to use for selecting a sample. Alternative weights have been proposed that make use of the information contained in the covariate and/or surrogate variable distributions to select samples that are more informative for the estimating model of interest. These include weights developed by Zhang, Ning, and Ruppert (2019) for the setting where only the covariates $\boldsymbol{x}_i$ are observed,

$$\pi_{2,yMISS} = \frac{\sqrt{p_i(\hat{\boldsymbol{\beta}})\left\{1 - p_i(\hat{\boldsymbol{\beta}})\right\}}||\boldsymbol{M}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_i||}{\sum_{j=1}^{n}\sqrt{p_j(\hat{\boldsymbol{\beta}})\left\{1 - p_j(\hat{\boldsymbol{\beta}})\right\}}||\boldsymbol{M}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_j||}$$

where $p_i(\hat{\boldsymbol{\beta}}) = \exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})/\left\{1 + \exp(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})\right\}$ and $\boldsymbol{M}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} p_i(\hat{\boldsymbol{\beta}})\left\{1 - p_i(\hat{\boldsymbol{\beta}})\right\}\boldsymbol{x}_i\boldsymbol{x}_i^T$.

Importantly, as the weights are model-based, they require some knowledge on the model of interest to determine the relative informativeness of observations. Hence, the weights are functions of an estimate of $\boldsymbol{\beta}$. To accomplish this, a two-step approach is used, where a smaller, preliminary sample is randomly selected and validated to provide a pilot estimate of $\boldsymbol{\beta}$, denoted as $\hat{\boldsymbol{\beta}}$, which is then included in the calculation of the weights. $\pi_{2,yMISS}$ gives greater weight to individuals with fitted probabilities closer to 0.5, or for whom the current model is most uncertain about their outcomes. Thus more information is gained by uncovering their true outcomes compared to other individuals.

Weights have also recently been proposed by Marks-Anglin et al. (2021) that additionally make use of s_i. One set of weights, called *surrogate augmented weights (sAUG)*, have the form,

$$\pi_{2,sAUG} = \frac{\sqrt{\left\{p_i(\hat{\boldsymbol{\alpha}}) - 2p_i(\hat{\boldsymbol{\alpha}})p_i(\hat{\boldsymbol{\beta}}) + p_i(\hat{\boldsymbol{\beta}})^2\right\}}||\boldsymbol{M}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_i||}{\sum_{j=1}^{n}\sqrt{\left\{p_j(\hat{\boldsymbol{\alpha}}) - 2p_j(\hat{\boldsymbol{\alpha}})p_j(\hat{\boldsymbol{\beta}}) + p_j(\hat{\boldsymbol{\beta}})^2\right\}}||\boldsymbol{M}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_j||},$$

where $\boldsymbol{\alpha}$ is the vector of coefficients from model $\text{logit}\left\{p(y_i = 1|s_i, \boldsymbol{x}_i)\right\} = (s_i, \boldsymbol{x}_i)^T\boldsymbol{\alpha}$.These weights were shown to offer greater efficiency compared to $\pi_{2,yMISS}$ when s_i is moderately or highly correlated with y_i.

The other set of proposed weights, called *surrogate substitution weights (sSUB)*, are optimal (most effiicent) for the model of s_i regressed on $\boldsymbol{x}_i$, but offer efficiency gains relative to both $\pi_{2,sAUG}$ and $\pi_{2,yMISS}$ under high sensitivity and specificity of s_i, as well as in rare event settings, and can be implemented in a single step (i.e. a preliminary sample does not need to be validated in order to construct the weights). These weights have the form,

$$\pi_{2,sSUB} = \frac{|s_i - p_i(\hat{\boldsymbol{\gamma}})|||\boldsymbol{Q}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_i||}{\sum_{j=1}^{n}|s_j - p_j(\hat{\boldsymbol{\gamma}})|||\boldsymbol{Q}_{\boldsymbol{x}}^{-1}\boldsymbol{x}_j||},$$

where $\boldsymbol{\gamma}$ is vector coefficient in the model $\text{logit}\{p(s_i = 1|\boldsymbol{x}_i)\} = \boldsymbol{x}_i^T\boldsymbol{\gamma}$ and $\boldsymbol{Q}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} p_i(\hat{\boldsymbol{\gamma}})(1 - p_i(\hat{\boldsymbol{\gamma}}))\boldsymbol{x}_i\boldsymbol{x}_i^T$.

We seek to explore how selection bias in phase 1 of an EHR-based study can impact inference made using the validation sample that is constructed using the aforementioned weights. Specifically, the validation sampling weights ($\pi_{2,yMISS}$, $\pi_{2,sAUG}$, $\pi_{2,sSUB}$) depend on pilot estimates (for example $\hat{p}(y_i|\boldsymbol{x}_i)$, or $\hat{p}(y_i|s_i, \boldsymbol{x}_i)$) obtained using some or all of the EHR study cohort. However, in the presence of selection bias, the disease model in the EHR cohort does not equate to the disease model in the target population, and so the constructed weights may actually be efficient for an alternative model (the disease model for those selected into the EHR cohort (model 4.3)), and not the true model of interest (model 4.1).

Beesley and Mukherjee (2020) show that in the presence of patient selection mechanisms, the model fit to the EHR cohort can be expressed as

$$\text{logit}\{P(y_i = 1|\boldsymbol{x}_i, R_{i,1} = 1)\} = \boldsymbol{x}_i\boldsymbol{\beta} + \frac{P(R_{i,1} = 1|y_i = 1, \boldsymbol{x}_i)}{P(R_{i,1} = 1|y_i = 0, \boldsymbol{x}_i)}. \tag{4.7}$$

Thus, bias in estimating $\boldsymbol{\beta}$ can occur when the final term in (4.7) varies with $\boldsymbol{x}_i$, which may in turn lead to biased estimates of $p_i(\hat{\boldsymbol{\beta}})$ and other quantities that are required for the sampling weights. We therefore expect that sampling weights based on the biased model may not be most efficient for the model of interest.

4.2.3. Inverse Probability of Selection Weighting

If the phase 1 selection weights $\pi_1(y_i, \boldsymbol{a}_i, \boldsymbol{\eta})$ are known, then unbiased pilot estimates ($\hat{\boldsymbol{\beta}}$) for the target model of interest can be obtained via inverse-probability weighted (IPW) estimation using the

following weighted log-likelihood,

$$\sum_{i=1}^{r_1} \frac{1}{\pi_1(y_i, \boldsymbol{a}_i, \boldsymbol{\eta})} \left[y_i \log \{p_i(\boldsymbol{\beta})\} + (1 - y_i) \log \{1 - p_i(\boldsymbol{\beta})\} \right] \tag{4.8}$$

Other pilot estimates needed for constructing the phase 2 weights (eg. $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\gamma}}$) can be similarly obtained through IPW estimation. These pilot estimates are then incorporated into $\pi_2(\boldsymbol{x}_i, s_i, \boldsymbol{\phi})$. This would ensure that observations sampled using the phase 2 weights are indeed informative for the model of interest (model 4.1), rather than a biased model (model 4.3). Furthermore, since $\pi_2(\boldsymbol{x}_i, s_i, \boldsymbol{\phi})$ are computed and known to the investigator, an unbiased estimate of $\boldsymbol{\beta}$ in the final sample can be obtained with the following weighted log-likelihood,

$$\sum_{i=1}^{r_1} \frac{1}{\pi_1(y_i, \boldsymbol{a}_i, \boldsymbol{\eta}) \pi_2(\boldsymbol{x}_i, s_i, \boldsymbol{\phi})} \left[y_i \log \{p_i(\boldsymbol{\beta})\} + (1 - y_i) \log \{1 - p_i(\boldsymbol{\beta})\} \right] \tag{4.9}$$

In practice, the phase 1 probabilities for selection into the EHR cohort from the target population ($\pi_1(y_i, \boldsymbol{a}_i, \boldsymbol{\eta})$) may not be known. Beesley and Mukherjee (2020) proposed methods to approximate the selection weights in the presence of external data from the target population. Specifically, if data on an external sample from the target population is available (for which the external sample selection process is known), and data on $\boldsymbol{a}_i$ and y_i are available,

$$P(R_{i,1} = 1 | \boldsymbol{a}_i, y_i) = P(R_{i,ext} | \boldsymbol{a}_i, y_i) \frac{p_{11} + p_{10}}{p_{11} + p_{01}}, \tag{4.10}$$

where $p_{jk} = P(R_{i,1} = j, R_{i,ext} = k | \boldsymbol{a}_i, y_i, R_{i,all} = 1)$, $R_{i,ext} = 1$ indicates inclusion in the external dataset, and $R_{i,all}$ indicates inclusion in either the internal EHR dataset or the external dataset. If the population size is large enough such that there is negligible overlap between the EHR cohort and the external dataset, the weights can be approximated by,

$$P(R_{i,1} = 1 | \boldsymbol{a}_i, y_i) \approx P(R_{i,ext} | \boldsymbol{a}_i, y_i) \frac{P(R_{i,1} = 1 | \boldsymbol{a}_i, y_i, R_{i,all} = 1)}{1 - P(R_{i,1} = 1 | \boldsymbol{a}_i, y_i, R_{i,all} = 1)}. \tag{4.11}$$

In the absence of an external dataset, summary statistics from the target population can be used to construct post-stratification weights taking the form

$$w_i \propto \frac{f(y_i, \boldsymbol{a}_i)}{f(y_i, \boldsymbol{a}_i | R_{i,1} = 1)} \tag{4.12}$$

where $f(y_i, \boldsymbol{a}_i)$ refers to the joint distribution of y_i and $\boldsymbol{a_i}$ in the target population.

Algorithm 1: Sampling for outcome validation accounting for selection bias

1. **Step One**: r_1 observations are selected with weights equal to $1/n$. For selected observations, chart review is performed to uncover the true outcome, y, and the sampled data points denoted as $\boldsymbol{O}_i^* = (\boldsymbol{x}_i^*, s_i^*, y_i^*, \pi_i^*)$, for $i = 1, \ldots, r_1$. Use pilot sample and external data (or summary statistics) from target population to estimate phase one weights, π_{1i} using (4.10), (4.11) or (4.12). If only a_i is needed to model phase 1 selection, weights can be calculated using full EHR cohort.
2. **Step Two**: Using pilot sample and estimated π_{1i}, preliminary estimates of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and/or $\boldsymbol{\gamma}$ are calculated accounting for phase 1 selection bias using the weighted log-likehood in (4.8).
3. **Step Three**: Use the pilot parameter estimates from step 2 to construct one of the more informative weights, $\pi_{2i} \in \{\pi_{i,yMISS}, \pi_{i,sAUG}, \pi_{i,sSUB}\}$, which is then used to sample r_2 individuals in the validation sample for further chart review. Estimation of $\boldsymbol{\beta}$ proceeds using the weighted log-likelihood in (4.9).
4. **Step Four**: Denote the pilot estimate for $\boldsymbol{\beta}$ from step 2 as $\tilde{\boldsymbol{\beta}}_1$, and the estimate from step 3 as $\tilde{\boldsymbol{\beta}}_2$. Use inverse variance weighting to combine the estimates:

$$\tilde{\boldsymbol{\beta}} = \left(\tilde{\boldsymbol{V}}_1^{-1} + \tilde{\boldsymbol{V}}_2^{-1}\right)^{-1}\left(\tilde{\boldsymbol{V}}_1^{-1}\tilde{\boldsymbol{\beta}}_1 + \tilde{\boldsymbol{V}}_2^{-1}\tilde{\boldsymbol{\beta}}_2\right), \qquad \tilde{\boldsymbol{V}} = \left(\tilde{\boldsymbol{V}}_1^{-1} + \tilde{\boldsymbol{V}}_2^{-1}\right)^{-1}$$

$$\tilde{\boldsymbol{V}}_1 = \tilde{\boldsymbol{M}}_{x,1}^{-1}\left\{\frac{1}{r_1^2}\sum_{i=1}^{r_1}\frac{1}{\pi_{1i}^2}\left(y_i - p_i(\tilde{\boldsymbol{\beta}}_1)\right)^2\boldsymbol{x}_i\boldsymbol{x}_i^T\right\}\tilde{\boldsymbol{M}}_{x,1}^{-1}$$

$$\tilde{\boldsymbol{V}}_2 = \tilde{\boldsymbol{M}}_{x,2}^{-1}\left\{\frac{1}{n^2 r_2^2}\sum_{i=1}^{r_2}\frac{1}{\pi_{1i}^2\pi_{2i}^2}\left(y_i - p_i(\tilde{\boldsymbol{\beta}}_2)\right)^2\boldsymbol{x}_i\boldsymbol{x}_i^T\right\}\tilde{\boldsymbol{M}}_{x,2}^{-1}$$

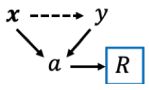
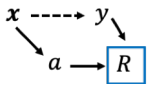
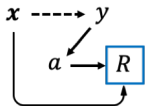
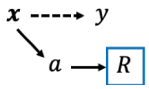
$$\tilde{\boldsymbol{M}}_{x,1} = \frac{1}{r_1}\sum_{i=1}^{r_1}\frac{1}{\pi_{1i}}p_i(\tilde{\boldsymbol{\beta}}_1)(1 - p_i(\tilde{\boldsymbol{\beta}}_1))\boldsymbol{x}_i\boldsymbol{x}_i^T \quad \tilde{\boldsymbol{M}}_{x,2} = \frac{1}{nr_2}\sum_{i=1}^{r_2}\frac{1}{\pi_{1i}\pi_{2i}}p_i(\tilde{\boldsymbol{\beta}}_2)(1 - p_i(\tilde{\boldsymbol{\beta}}_2))\boldsymbol{x}_i\boldsymbol{x}_i^T$$

Note that since y_i is unobserved in the internal EHR cohort, we cannot use the full dataset to estimate $P(R_{i,1} = 1|\boldsymbol{a}_i, y_i, R_{i,all} = 1)$ or $f(y_i, \boldsymbol{a}_i|R_{i,1} = 1)$. However, if a pilot sample from the internal EHR dataset is randomly selected and validated, such that y_i is observed for a small subset of individuals (as is already done in two-step validation sampling procedures), then the pilot sample data can be used to estimate this quantity. We therefore propose the procedure in Algorithm 1 for sampling and analyzing a validation sample while accounting for selection bias.

4.3. Simulation Study

We simulate a variety of selection settings to study the performance of the validation study sampling approaches with and without the IPW and poststratification methods outlined in Section 4.2.3. To account for selection bias, we study the use of the true selection weights, selection weights estimated using an external dataset, and post-stratification weights. For each of $M = 250$ simulates, we generate the target population of size $N = 100,000$, with true disease model coefficients of $\boldsymbol{\beta} = (0.5, 0.5)$ for $\boldsymbol{x}_i \in \Re^2$. Additionally, a continuous variable $a_i \notin \boldsymbol{x}_i$ influences selection into the EHR cohort, according to the selection settings described in Table 4.1. In some settings we induce association between y_i and $\boldsymbol{a}_i$ by defining a latent variable $\boldsymbol{a}_{i,original}$, such that $\boldsymbol{a}_i = f(\boldsymbol{a}_{i,original}, y_i)$. Otherwise, $\boldsymbol{a}_i = \boldsymbol{a}_{i,original}$. From the target population, approximately $n = 25,000$ to $35,000$ individuals (depending on the selection mechanism) are selected into the EHR cohort.

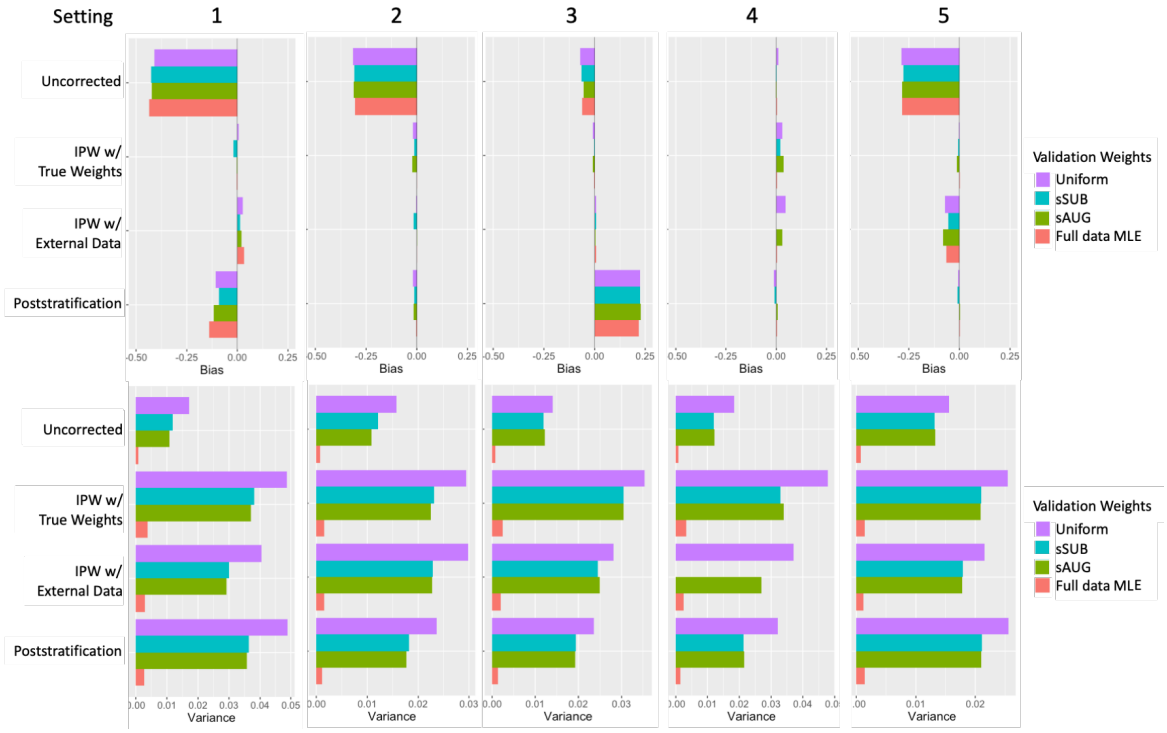
Table 4.1: Selection settings for simulation with corresponding directed acyclic graphs (DAGs). Selection probabilities depend on (1) a_i only, where a_i is correlated with both y_i and $\boldsymbol{x}_i$; (2) a_i and y_i, where a_i is correlated with $\boldsymbol{x}_i$ only; (3) a_i and x_i, where a_i is correlated with y_i only; (4) a_i only, where a_i is correlated with x_i only; (5) a_i only, where a_i is a modifier for the effect of x_{i1} on y_i.

Setting	logit$\{P(R_{i,1} = 1)\}$	a_i	cov$(a_{i,original}, x_{i,1})$	Causal DAG
1	$-0.2 - 2a_i$	$a_{i,original} + 2y_i$	0.7	
2	$-0.2 - 1.5a_i - 1.5y_i$	$a_{i,original}$	0.7	
3	$-0.2 - 1.5a_i - 1x_{i1}$	$a_{i,original} + 1y_i$	0	
4	$-1 - 2a_i$	$a_{i,original}$	0.7	
5	$-0.2 - 3a_i$	$a_{i,original}$	0	effect modification for x_{i1} by a_i

For outcome misclassification in the EHR cohort, sensitivity and specificity of the surrogate out-

come are set to 90% and 80% respectively. Outcome validation proceeds with a pilot sample of $r_1 = 500$ individuals, and a more informative sample of $r_2 = 1500$ individuals, on which the final model is fit. The validation sampling methods compared include uniform sampling, surrogate augmented sampling, and surrogate substitution sampling. For calculating poststratification weights, a_i is binned into ranges to form a categorical variable. We also include estimates using the full EHR cohort. Empirical bias and variance are calculcated over the M replicates, comparing the validation sample-based estimates of β_1 to the true values.

Figure 4.2: Empirical bias and variance of β_1 over $M = 250$ simulates in settings 1-5, as outlined in Table 4.1.



In Figure 4.2 we observe that the uncorrected analysis results in bias in estimating β_1 for all settings, with the exception of setting 4, in which selection only depends on the exposure variable. For settings 1, 2, 3, and 5, IPW using either the true selection weights or weights estimated with the aid of external data resulted in substantial bias reduction, though some bias remains for IPW w/ external data if the bias is due to selection on an effect modifier. Poststratification did not perform as well as the other two forms of adjustment, likely due to the categorizing of continuous a_i to calculate the weights, which does not fully capture the distributional differences in a_i between the

target population and the study cohort. Poststratification actually produced greater bias compared to the uncorrected approach for setting 3.

Weighted approaches to correct for selection bias resulted in greater variance in the estimates in all settings. Inflated variance is a known limitation of IPW, but the increase in uncertainty compared to the uncorrected estimates is small relative to the bias reduction for most approaches.

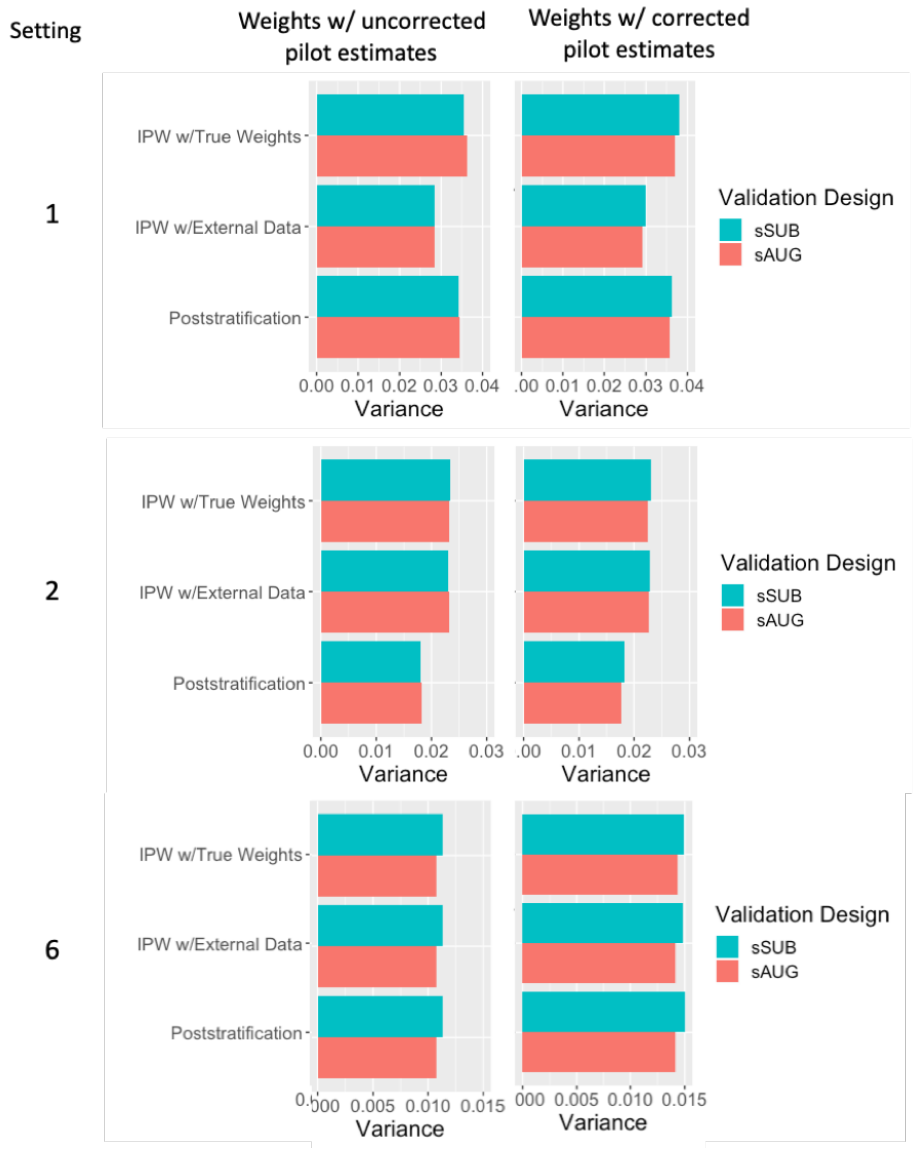
We performed further simulations comparing the performance of the validation sampling weights when the step 1 pilot estimates were corrected for selection bias, compared to using weights that are constructed with biased pilot estimates. The results are shown in Figure 4.3 for simulation settings 1, 2, and a 3rd setting where selection is directly a function of y_i and $x_{i,1}$ (denoted here as setting 6). We find that constructing the sampling weights using the uncorrected, biased pilot estimates leads to similar or even better precision in comparison to the use of weights based on corrected pilot estimates. This finding is may be due to the inflation of variance cause by IPW in the pilot sample estimates.

4.4. Colon Cancer Recurrence Study

In this section we seek to study the association between the stage of a primary colon cancer and the risk of colon cancer recurrence within 5 years using an EHR dataset (Chubak et al., 2018). The dataset includes 1,063 patients from an integrated health system, Kaiser Permanente Washington (KPWA), who were diagnosed with stage I-IIIA malignant colorectal adenocarcinomas between 1995 and 2014. This dataset is especially useful as both the gold-standard, validated date of index colon cancer recurrence, as well as an algorithmically derived phenotype developed using structured administrative and registry data (Hassett et al., 2017) are available, which we denote the potentially misclassified outcome s_i. In practice we expect only a subset of patients would have outcomes available.

Kaiser Permanente is the largest nonprofit integrated health care delivery system in the US. However, populations enrolled in KPWA are primarily of white race, which is reflective of the population in Washington, but is not representative of the national U.S population. KPWA also includes insured individuals only, the majority of whom are employed. Furthermore, for the purpose of illustrating our approach, we treat 5-year cancer recurrence as a binary outcome, and exclude patients who were

Figure 4.3: Mean variance of $\tilde{\beta}_1$ over 500 replicates in settings 1, 2, and 6 (where selection depends directly on both y_i and $x_{i,1}$). In left panel, sampling weights constructed with biased pilot estimates. In right panel, sampling weights constructed with pilot estimates corrected for selection bias

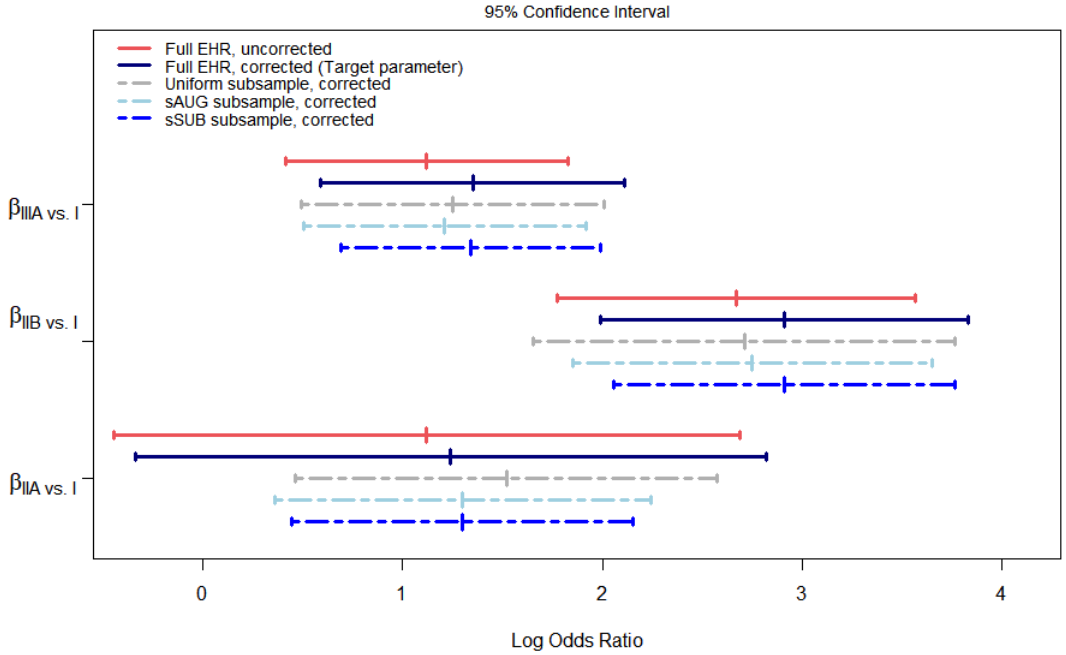


lost to follow-up (primarily due to disenrollment from Kaiser's health plan) prior to the 5-year mark, reducing the final cohort to 771 patients. Disenrollment occured disproportionately among younger individuals in this cohort, thus the age distribution is also not representative. If race and/or age are effect modifiers of the association between colon cancer stage and recurrence, the selection process for the final EHR cohort could therefore result in limited transportability of the estimated treatment effect to the general U.S. population. To account for this nonrepresentativeness, we used incidence rate data from the Surveillance, Epidemiology, and End Results (SEER) Program at the National Cancer Institute, combined with Census population data from 2004-2014, to estimate the joint distribution of race and age among patients in the US with stages 1-IIIA colon cancer. Using these summary statistics, we calculated poststratification weights to incorporate in our validation study analysis.

For the subsample-based association study, a first stage sample of 300 patients was randomly selected, followed by an additional 200 patients with more informative weights (using surrogate substitution or surrogate augmented weights) or uniform weights. Due to the relatively small sample sizes for a rare outcome ($\sim 7\%$), a Firth adjustment (Firth, 1993; Rader et al., 2017) was used in logistic regression for all analyses. We repeated the validation sampling procedure 500 times and calculated mean parameter estimates and standard errors.

In total, 57 patients were diagnosed with recurrent colon cancer within 5 years, while the phenotyping algorithm identified 160 patients as having the outcome of interest. Overall sensitivity and specificity are equal to 91% and 85% respectively. The results displayed in Figure 4.4 show that the estimates accounting for population selection based on race and age are slightly farther from the null compared to the naive analysis. We observe similar correction in the subsample-based estimates, with narrower confidence intervals attained using the surrogate-assisted weights as compared to the uniform weights. The subsamples constructed using the surrogate-substitution weights yielded estimates closest to the corrected, full-data based analysis, with narrower confidence intervals. The greater precision in the subsample-based estimates compared to the full-data estimates could be attributed to the study designs allowing for sampling with replacement, combined with the low incidence of both the event and exposure variables. When sampling with replacement, the efficient sampling weights can select more informative observations multiple times, improving precision in the final estimates.

Figure 4.4: Log-odds ratio estimates and 95% confidence intervals for the effect of SEER stage of index colon cancer (relative to stage I) on colon cancer recurrence within 5 years, adjusting for year of index colon cancer diagnosis



4.5. Discussion

In the presence of outcome misclassification, informative sampling designs to guide patient selection for chart review are useful for obtaining efficient, internally valid association parameter estimates in measurement constrained settings. However, when extending inference from EHR-based studies to larger or different populations, external validity is equally important. Nonrepresentativeness of EHR databases due to selection mechanisms influencing which and how often patients interact with healthcare systems can lead to selection bias and limit the generalizability and transportability of EHR cohort-based results. This dual consideration of outcome misclassification and selection bias is a complex issue that has only been recently explored (Beesley and Mukherjee, 2020; Zhang et al., 2020). In this work we have illustrated the application of inverse probability of selection weighting to improve the external validity of findings from subsamples constructed using efficient sampling weights.

We have shown that failure to account for patient selection can result in biased model estimates in association studies using EHR data, even when the gold-standard outcomes are available. Bias

is observed to primarily occur when the selection mechanism depends on both the outcome of interest and the exposure variable, either directly or through a related auxilliary variable. Our simulaions show that selection based on the exposure variable alone does not result in bias, but rather bias may be induced if additional adjustment is made through IPW in this setting. However, we do consider and illustrate the potential for incorrect inference on overall treatment effect when effect modifiers are distributed differently in the EHR cohort and the target population. This form of nonrepresentativeness is less often considered in the context of selection bias, but is an important consideration for the external validity of study results.

Our work assumes that individual-level external data or summary statistics for the joint distribution of selection variables in the target population are available. This may not always be feasible. If the joint distribution of all variables influencing selection are not available, but marginal distributions are obtainable, then raking (also known as iterative proportional fitting) can be considered for approximating the poststratification weights (Deville, Särndal, and Sautory, 1993). Raking offers the advantage of incorporating more variables in the weighting process, and is easier to implement with small samples. If marginal frequencies are also unavailable, sensitivity analysis can be performed under a range of possible selection mechanisms.

Importantly, while the additional weighting by selection probabilities produces substantial bias reduction in most of the settings studied, this comes at the cost of inflated variance. We found in simulations that correcting for selection bias resulted in greater variance relative to an uncorrected analysis. Furthermore, accounting for selection bias in the preliminary step of the validation study design did not lead to precision gains compared to using uncorrected, biased pilot estimates to construct the validation sampling weights. Indeed, we observed in one setting that using the uncorrected pilot estimates to construct the validation sampling weights actually produced better precision. Future applications should consider using more robust variance estimators when applying IPW to validation studies.

CHAPTER 5

DISCUSSION

The reproducibility of evidence is a hallmark of good science, and crucial to the advancement of medicine. Systemic sources of bias threaten both the internal and external validity of research findings, undermining the goals of scientific inquiry. In this dissertation we studied potential sources of systemic bias that can impact evidence-based decision making in healthcare settings. These include bias from selective publication of experimental studies, bias in observational studies from differential misclassification of EHR-derived phenotypes, and bias due to patient selection in EHR-based cohorts. It is important to note that these are all forms of selective observation, which distorts the evidence-base informing treatment protocols, standard of care and health policy. We proposed novel statistical methodologies to address these sources of bias, including a new estimation procedure for fitting selection models accounting for publication bias in meta-analyses, cost-effective study designs for validating EHR outcomes in measurement-constrained settings, and algorithms to improve the generalizability of association-study results from subsamples selected for outcome validation. These approaches were studied and validated through simulation studies as well as the use of datasets with unique features to enable performance evaluation, such as the availability of both published and unpublished studies in a dataset for meta-analysis, and gold standard outcomes along with EHR-derived phenotypes in a real-world EHR dataset.

Our EM algorithm for fitting Copas' selection model in the network meta-analysis (NMA) framework was shown to provide substantial bias reduction in small sample settings relative to model complexity. This approach also enables us to assume unstructured heterogeneity in a setting where model flexibility is important. We acknowledge the limitations associated with EM algorithms in general, primarily the sensitvity of results to the initial values used, and propose intuitive guidelines for selecting starting values. Future work in this area will aim to expand this model to larger networks, which may require imposing constraints on our maximally flexible working model, while also developig innovative measures for assessing publication bias in the network setting (for example, visualizing indirect effects).

The validation sampling designs we proposed in Aims 2 and 3 offer a means for reducing the cost

and time of validation studies using EHRs, by leveraging information contained in the potentially misclassified EHR-derived phenotypes to prioritize the most informative patients for sample selection. We illustrated the advantages and disadvantages of our two candidate sampling schemes (surrogate substitution and surrogate augmented sampling), particularly through the lens of robustness and statistical efficiency, and outlined settings for which each one is most appropriate. An important limitation is that these designs require that the model of interest is known *a priori*, and extensions are needed for model building and model selection.

We note that evidence-based medicine is not only concerned with evidence generation, but rather it integrates the best available evidence with clinical expertise along with a patient's profile and preferences in order to make optimal decisions for care and treatment. Recent emphasis on personalized medicine and decision-making (Greenhalgh, Howick, and Maskrey, 2014) has led to the consideration of patient-specific risk profiles based on biological, serological and social factors in clinical decision making (see Figure 5.1 by (Makam and Nguyen, 2017)). Furthermore, advanced electronic clinical decision support systems (CDSSs) that integrate with EHR systems and leverage patient characteristics, individual risk assessments based on EHRs, and up-to-date guidelines (including protocols based on results from MAs) are increasingly used at point-of-care to aid in decision-making and improve clinical outcomes (Classen et al., 1991; Gianfrancesco et al., 2018; Kilbridge et al., 2006; Manaktala and Claypool, 2017; Rudin et al., 2019; Sim et al., 2001). In summary, evidence on the relative efficacy and safety of treatments are often considered in conjunction with the patient's risk profile for the outcome of interest or adverse events to determine the optimal treatment selection.

A natural extension of the work in this dissertation would therefore be to address sources of bias in patient risk assessment for clinical decision-making. Notably, diagnostic models that predict individual disease status and prognostic models that predict a future adverse event, hospital readmission, or discharge can be subject to algorithmic bias if the training dataset is not a representative sample of the population of interest (Gianfrancesco et al., 2018). In particular, a lack of respresentation of underserved groups (defined by genetic-based ancestry, sex, ethnicity, socioeconomic status, insurance status, geographic origin) in biobanks and EHR databases is a known issue often resulting in a biased evidence-base for personalized medicine. The sampling algorithms studied in Aims 2 and 3 can be extended to the framework of patient recruitment for prediction modeling, also known

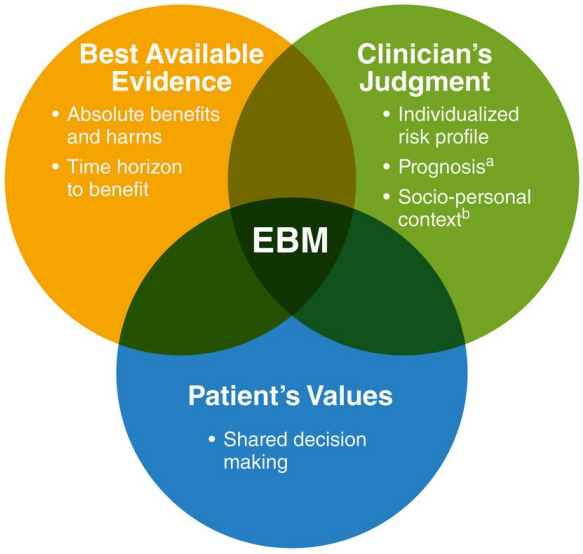


Figure 5.1: Evidence-based medicine framework for clinical decision making. (Makam and Nguyen, 2017)

as active learning. An oversampling of underrepresented groups in precision medicine studies (as opposed to uniform sampling) can lead to a more robust database for building predictive models. This future work will also seek to incorporate fairness criteria to ensure clinical decisions for patients are made equitably and do not perpetuate health disparities (Ferryman and Pitcan, 2018).

APPENDIX A

CHAPTER 2 APPENDICES

A.1. Sensitivity of results to mispecification of heterogeneity structure

We performed additional simulations to investigate the performance of Copas' model when formulated with fully structured, common between-study heterogeneity (i.e. $\tau_1 = \tau_2 = \tau_3 = \tau_4$), as opposed to the true data generating mechanism which is unstructured, allowing for different levels of heterogeneity by study design. These resutls are shown in Table A.1.

Table A.1: Empirical bias of parameter estimates when assuming (a) different between-study heterogeneities (correctly specified model) vs. (b) common heterogeneity (misspecified model) with τ initialized at the aveage value. NPR = 50%

		Naïve NMA				EMBRACE				
n_d	PAR	MEAN BIAS	ESE	MBSE	REL BIAS	MEAN BIAS	ESE	MBSE	REL BIAS	BIAS RED
(a) Different heterogeneities: $\boldsymbol{\tau}^2 = (\tau_1^2, \tau_2^2, \tau_3^2, \tau_4^2)$										
10	μ^{AC}	9.0	19.1	16.1	11.3	-0.2	21.7	22.1	-0.3	102%
	μ^{BC}	15.7	19.1	17.9	31.4	-0.4	26.8	23.6	-0.8	102%
	μ^{BA}	6.7	21.5	18.9	-22.3	-0.2	27.3	27.3	0.6	103%
						NCR=2.4%				
(b) Common heterogeneity: $\tau_1^2 = \tau_2^2 = \tau_3^2 = \tau_4^2$; Initial value = mean($\hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3, \hat{\tau}_4$) $\approx$ 0.835										
10	μ^{AC}	10.1	19.0	18.2	12.6	0.5	25.7	27.2	0.7	95%
	μ^{BC}	15.9	18.2	18.2	31.8	-2.2	27.6	28.0	-4.5	114%
	μ^{BA}	5.8	20.9	20.0	-19.4	-2.8	30.4	32.1	9.2	148%
		$\hat{\tau}$ = 0.840				$\hat{\tau}$ = 0.837				
						NCR=0.0%				

ESE: empirical standard error; MBSE: model based standard error using the first 3×3 submatrix of the observed information matrix; REL BIAS: relative empirical bias; BIAS RED: percent bias reduction relative to naïve estimates; NCR: non-convergence rate

BIAS, ESE, MBSE entries multiplied by 100

A.2. Sensitivity of results to initial values of EM algorithm

In this section, we assess algorithm performance when starting values for the selection parameters are farther from the truth (eg. $+\kappa \sim Uniform(-0.5, 0.5)$ for α_d and β_d, or between 0 and 1 for $\{\rho_1, \rho_2, \rho_3, \rho_4^{AC}, \rho_5^{BC}\}$. The results are shown in Table A.2.

Table A.2: Results of EMBRACE when (a) initial values for selection parameters are close to the true values for all parameters, (b) far from the truth for α_d and ρ_d and (c) far from the truth for all parameters. NPR = 50%

n_d	PAR	MEAN BIAS	ESE	MBSE	REL BIAS	BIAS RED	NCR
(a) starting values = $\{\alpha_d, \beta_d, \rho_1, \rho_2, \rho_3, \rho_4^{AC}, \rho_4^{BC}\} + \kappa \sim Unif(-0.1, 0.1)$							
10	μ^{AC}	-0.2	21.7	22.1	-0.3	102%	2.4%
	μ^{BC}	-0.4	26.8	23.6	-0.8	102%	
	μ^{BA}	-0.2	27.3	27.3	0.6	103%	
(b) starting values = $\beta_d + \kappa \sim Unif(-0.1, 0.1)$, $\alpha_d + \kappa \sim Unif(-0.5, 0.5)$ $\{\rho_1, \rho_2, \rho_3, \rho_4^{AC}, \rho_4^{BC}\} \sim Unif(0, 1)$							
10	μ^{AC}	1.1	28.3	23.6	1.3	88%	5%
	μ^{BC}	0.3	36.6	26.9	0.6	98%	
	μ^{BA}	-0.7	40.2	30.3	2.4	111%	
(c) starting values = $\{\alpha_d, \beta_d\} + \kappa \sim Unif(-0.5, 0.5)$ $\{\rho_1, \rho_2, \rho_3, \rho_4^{AC}, \rho_4^{BC}\} \sim Unif(0, 1)$							
10	μ^{AC}	-1.5	48.4	19.1	-1.9	117%	29.4%
	μ^{BC}	-0.9	51.2	22.1	-1.7	105%	
	μ^{BA}	0.6	64.0	26.2	-2.1	90%	

ESE: empirical standard error; MBSE: model based standard error using the first 3×3 submatrix of the observed information matrix; REL BIAS: relative empirical bias; BIAS RED: percent bias reduction relative to naïve estimates; NCR: non-convergence rate

BIAS, ESE, MBSE entries multiplied by 100

APPENDIX B

CHAPTER 3 APPENDICES

B.1. Derivation of Asymptotic Properties of $\tilde{\boldsymbol{\beta}}$

Recall from Section 2.2 that we wish to estimate $\tilde{\boldsymbol{\beta}}$ as the solution to the reweighted score equation

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

where $\{\pi_i^*\}_{i=1}^r$ represent the sampling probabilities of observations $i = 1, ..., r$ selected in the sample. By Taylor's theorem, it can be shown that

$$\begin{aligned}
0 = \frac{\dot{\ell}^*(\tilde{\boldsymbol{\beta}})}{n} &= \frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n} + \frac{1}{n}\frac{\delta\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{\delta\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}) + O_{P|\mathscr{F}}(||\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}||^2) \\
&= \frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n} + \frac{1}{n}\frac{\delta\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{\delta\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}) + O_{P|\mathscr{F}}(r^{-1})
\end{aligned}$$

Then the influence function for $\tilde{\boldsymbol{\beta}}$ has the form

$$(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}) = \tilde{\boldsymbol{M}}_X^{-1}\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n} + O_{P|\mathscr{F}}(r^{-1}) \tag{B.1}$$

where $\tilde{\boldsymbol{M}}_X = \frac{1}{n}\frac{\delta\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{\delta\boldsymbol{\beta}^T} = \frac{1}{nr}\sum_{i=1}^{r}\frac{p_i^*(\hat{\boldsymbol{\beta}}_{MLE})(1 - p_i^*(\hat{\boldsymbol{\beta}}_{MLE}))\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}}{\pi_i^*}$

We also have that $E[\tilde{\boldsymbol{M}}_X|\mathscr{F}] = \boldsymbol{M}_X$ and for any component $\tilde{\boldsymbol{M}}_X^{j_1,j_2}$ of $\boldsymbol{M}_X$, where $0 \leq j_1, j_2 \leq d$,

$$\begin{aligned}
\text{Var}[\tilde{\boldsymbol{M}}_X^{j_1,j_2}|\mathscr{F}] &= \frac{1}{r}\sum_{i=1}^{n}\pi_i\left\{\frac{p_i(\hat{\boldsymbol{\beta}}_{MLE})(1 - p_i(\hat{\boldsymbol{\beta}}_{MLE}))x_{ij_1}x_{ij_2}^T}{n\pi_i} - \boldsymbol{M}_X^{j_1,j_2}\right\}^2 \\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\pi_i\frac{p_i(\hat{\boldsymbol{\beta}}_{MLE})(1 - p_i(\hat{\boldsymbol{\beta}}_{MLE}))^2(x_{ij_2}x_{ij_1}^T)^2}{\pi_i^2} - \frac{1}{r}\sum_{i=1}^{n}\pi_i\Big(\boldsymbol{M}_X^{j_1,j_2}\Big)^2 \\
&= \frac{1}{rn^2}\sum_{i=1}^{n}\frac{p_i(\hat{\boldsymbol{\beta}}_{MLE})(1 - p_i(\hat{\boldsymbol{\beta}}_{MLE}))^2(x_{ij_1}x_{ij_2}^T)^2}{\pi_i} - \frac{1}{r}\Big(\boldsymbol{M}_X^{j_1,j_2}\Big)^2 \\
&\leq \frac{1}{16rn^2}\sum_{i=1}^{n}\frac{||\boldsymbol{x}_i||^4}{\pi_i} - \frac{1}{r}\Big(\boldsymbol{M}_X^{j_1,j_2}\Big)^2 \\
&\quad \text{(since } 0 \leq p_i(\hat{\boldsymbol{\beta}}_{MLE})(1 - p_i(\hat{\boldsymbol{\beta}}_{MLE})) \leq 0.25\text{)} \\
&= O_P(r^{-1}) \qquad \text{(due to the assumption that } \frac{1}{n^2}\sum_{i=1}^{n}\frac{||\boldsymbol{x}_i||^4}{\pi_i} = O_P(1)\text{)}
\end{aligned}$$

Thus by Chebyshev's inequality, $\tilde{\boldsymbol{M}}_X \xrightarrow{P|\mathscr{F}} \boldsymbol{M}_X$

We can therefore rewrite

$$\begin{aligned}
(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}) &= \tilde{\boldsymbol{M}}_X^{-1} \frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n} + O_{P|\mathscr{F}}(r^{-1}) \\
&= \tilde{\boldsymbol{M}}_X^{-1} \frac{1}{r} \sum_{i=1}^{r} \psi_i + O_{P|\mathscr{F}}(r^{-1}) \\
&= \tilde{\boldsymbol{M}}_X^{-1} \frac{1}{r} \sum_{i=1}^{r} \psi_i + O_{P|\mathscr{F}}(r^{-1}) - (\boldsymbol{M}_X^{-1} - \boldsymbol{M}_X^{-1}) \frac{1}{r} \sum_{i=1}^{r} \psi_i + O_{P|\mathscr{F}}(r^{-1}) \\
&= \boldsymbol{M}_X^{-1} \frac{1}{r} \sum_{i=1}^{r} \psi_i - (\tilde{\boldsymbol{M}}_X^{-1} - \boldsymbol{M}_X^{-1}) \frac{1}{r} \sum_{i=1}^{r} \psi_i + O_{P|\mathscr{F}}(r^{-1}) \\
&= \boldsymbol{M}_X^{-1} \frac{1}{r} \sum_{i=1}^{r} \psi_i + O_{P|\mathscr{F}}(r^{-1/2}) + O_{P|\mathscr{F}}(r^{-1}) \\
&= \boldsymbol{M}_X^{-1} \frac{1}{r} \sum_{i=1}^{r} \psi_i + O_{P|\mathscr{F}}(r^{-1/2})
\end{aligned} \tag{B.2}$$

Since $\dfrac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n} = \dfrac{1}{r} \sum_{i=1}^{r} \dfrac{(y_i^* - p_i^*(\tilde{\boldsymbol{\beta}}))\boldsymbol{x}_i^*}{n\pi_i^*} = \dfrac{1}{r} \sum_{i=1}^{r} \psi_i$, where $\psi_1, \psi_2, \ldots, \psi_r$ are i.i.d with mean 0 and variance equal to

$$\text{Var}(\psi_i|\mathscr{F}) = \frac{1}{n^2} \sum_{i=1}^{n} \frac{(y_i - p_i(\hat{\boldsymbol{\beta}}_{MLE}))^2 \boldsymbol{x}_i \boldsymbol{x}_i^T}{\pi_i} = O_P(1), \tag{B.3}$$

then by the central limit theorem

$$r^{1/2} \text{Var}(\psi_i|\mathscr{F})^{-1/2} \sum_{i=1}^{r} \psi_i \xrightarrow{D|\mathscr{F}} N(0, \boldsymbol{I}) \tag{B.4}$$

Combining (B.4) with (B.2) and using Slutsky's theorem, we have that

$$r^{1/2} \boldsymbol{V}^{-1/2} (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}) \xrightarrow{D|\mathscr{F}} N(0, \boldsymbol{I}) \tag{B.5}$$

where $\boldsymbol{V} = \boldsymbol{M}_X^{-1} \text{Var}[\psi_i|\mathscr{F}] \boldsymbol{M}_X^{-1}$

B.2. Derivation of Surrogate Augmented Weights (Proposition 1)

If y_i is known, one can proceed to derive the optimal sampling weights by minimizing the trace of $\boldsymbol{V}$, as is done in Wang, Zhu, and Ma (2018). However, if y_i is not known, we propose to first use the law of total variance as follows,

$$\begin{aligned}\mathsf{Var}\left(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{x}\right) =& \mathsf{E}\left\{\mathsf{Var}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s},\boldsymbol{y},\boldsymbol{x})|\boldsymbol{s},\boldsymbol{x}\right\} \;+\; \mathsf{E}\left[\mathsf{Var}\left\{\mathsf{E}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s},\boldsymbol{y},\boldsymbol{x}\right\}|\boldsymbol{s},\boldsymbol{x})|\boldsymbol{x}\right] \\ &+\; \mathsf{Var}\left\{\mathsf{E}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s},\boldsymbol{x})|\boldsymbol{x}\right\} \end{aligned} \tag{B.6}$$

From (B.1) we have that

$$\begin{aligned}
&\mathsf{E}\left\{\mathsf{Var}(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s},\boldsymbol{y},\boldsymbol{x})|\boldsymbol{s},\boldsymbol{x}\right\} = \mathsf{E}\left\{\mathsf{Var}(\boldsymbol{M}_X^{-1}\frac{1}{r}\sum_{i=1}^{r}\psi_i|\boldsymbol{s},\boldsymbol{y},\boldsymbol{x})|\boldsymbol{s},\boldsymbol{x}\right\} \\
&= \mathsf{E}\left[\frac{1}{r^2}\boldsymbol{M}_X^{-1}\sum_{i=1}^{r}\mathsf{Var}\left\{\frac{(y_i-p_i(\hat{\boldsymbol{\beta}}_{MLE}))\boldsymbol{x}_i}{n\pi_i}\Big|\boldsymbol{s},\boldsymbol{y},\boldsymbol{x}\right\}\boldsymbol{M}_X^{-1}\Big|\boldsymbol{s},\boldsymbol{x}\right] \\
&\overset{iid}{=} \mathsf{E}\left[\frac{1}{r}\boldsymbol{M}_X^{-1}\mathsf{Var}\left\{\frac{(y_i-p_i(\hat{\boldsymbol{\beta}}_{MLE}))\boldsymbol{x}_i}{n\pi_i}\Big|\boldsymbol{s},\boldsymbol{y},\boldsymbol{x}\right\}\boldsymbol{M}_X^{-1}\Big|\boldsymbol{s},\boldsymbol{x}\right] \\
&= \mathsf{E}\Bigg(\frac{1}{r}\boldsymbol{M}_X^{-1}\Bigg[E\left\{\frac{(y_i-p_i(\hat{\boldsymbol{\beta}}_{MLE}))^2\boldsymbol{x}_i\boldsymbol{x}_i^T}{n^2\pi_i^2}\Big|\boldsymbol{s},\boldsymbol{y},\boldsymbol{x}\right\} \\
&\quad - E\left\{\frac{(y_i-p_i(\hat{\boldsymbol{\beta}}_{MLE}))\boldsymbol{x}_i}{n\pi_i}\Big|\boldsymbol{s},\boldsymbol{y},\boldsymbol{x}\right\}^2\Bigg]\boldsymbol{M}_X^{-1}\Big|\boldsymbol{s},\boldsymbol{x}\Bigg) \\
&= \mathsf{E}\Bigg\{\frac{1}{n^2r}\boldsymbol{M}_X^{-1}\Bigg(\sum_{i=1}^{n}\pi_i\left\{\frac{(y_i-p_i(\hat{\boldsymbol{\beta}}_{MLE}))^2\boldsymbol{x}_i\boldsymbol{x}_i^T}{\pi_i^2}\right\} \\
&\quad - \left[\sum_{i=1}^{n}\pi_i\left\{\frac{(y_i-p_i(\hat{\boldsymbol{\beta}}_{MLE}))\boldsymbol{x}_i}{\pi_i}\right\}\right]^2\Bigg)\boldsymbol{M}_X^{-1}\Big|\boldsymbol{s},\boldsymbol{x}\Bigg\} \\
&= \mathsf{E}\left[\frac{1}{n^2r}\boldsymbol{M}_X^{-1}\left\{\sum_{i=1}^{n}\frac{(y_i-p_i(\hat{\boldsymbol{\beta}}_{MLE}))^2\boldsymbol{x}_i\boldsymbol{x}_i^T}{\pi_i}-\sum_{i=1}^{n}(y_i-p_i(\hat{\boldsymbol{\beta}}_{MLE}))^2\boldsymbol{x}_i\boldsymbol{x}_i^T\right\}\boldsymbol{M}_X^{-1}\Big|\boldsymbol{s},\boldsymbol{x}\right] \\
&= \frac{1}{n^2r}\boldsymbol{M}_X^{-1}\left[\sum_{i=1}^{n}\left\{p(y_i^2|s_i,\boldsymbol{x}_i)-2*p(y_i|s_i,\boldsymbol{x}_i)p_i(\hat{\boldsymbol{\beta}}_{MLE})+p_i(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}\boldsymbol{x}_i\boldsymbol{x}_i\Big(\frac{1}{\pi_i}-1\Big)\right]\boldsymbol{M}_X^{-1} \\
&\quad \text{since } p(y_i^2|s_i,\boldsymbol{x}_i)=p(y_i|s_i,\boldsymbol{x}_i) \\
&= \frac{1}{n^2r}\boldsymbol{M}_X^{-1}\left[\sum_{i=1}^{n}\left\{p_i(\hat{\boldsymbol{\alpha}}_{MLE})-2*p_i(\hat{\boldsymbol{\alpha}}_{MLE})p_i(\hat{\boldsymbol{\beta}}_{MLE})+p_i(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}\boldsymbol{x}_i\boldsymbol{x}_i\Big(\frac{1}{\pi_i}-1\Big)\right]\boldsymbol{M}_X^{-1}
\end{aligned} \tag{B.7}$$

where $\hat{\boldsymbol{\alpha}}_{MLE}$ is the estimated coefficient in the assumed working model for $p(y_i|s_i, \boldsymbol{x}_i)$. This would be estimated in step 1 using the first r_1 observations sampled.

Furthermore, we have

$$\begin{aligned}
&\mathsf{E}\left[\mathsf{Var}\left\{\mathsf{E}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{x}\right)|\boldsymbol{s}, \boldsymbol{x}\right\}|\boldsymbol{x}\right] = \mathsf{E}\left[\mathsf{Var}\left\{\mathsf{E}\left(\boldsymbol{M}_X^{-1}\frac{1}{r}\sum_{i=1}^{r}\psi_i|\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{x}\right)|\boldsymbol{s}, \boldsymbol{x}\right\}|\boldsymbol{x}\right] \\
&= \mathsf{E}\left(\mathsf{Var}\left[\boldsymbol{M}_X^{-1}\mathsf{E}\left\{\frac{(y_i - p_i(\hat{\boldsymbol{\beta}}_{MLE}))\boldsymbol{x}_i}{n\pi_i}\Big|\boldsymbol{s}, \boldsymbol{y}, \boldsymbol{x}\right\}\Big|\boldsymbol{s}, \boldsymbol{x}\right]\Big|\boldsymbol{x}\right) \\
&= \mathsf{E}\left[\mathsf{Var}\left\{\frac{1}{n}\boldsymbol{M}_X^{-1}\sum_{i=1}^{n}(y_i - p_i(\hat{\boldsymbol{\beta}}_{MLE}))\boldsymbol{x}_i\Big|\boldsymbol{s}, \boldsymbol{x}\right\}\Big|\boldsymbol{x}\right] \\
&= \frac{1}{n^2}\boldsymbol{M}_X^{-1}\sum_{i=1}^{n}p_i(\hat{\boldsymbol{\alpha}}_{MLE})(1 - p_i(\hat{\boldsymbol{\alpha}}_{MLE}))\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{M}_X^{-1}
\end{aligned} \tag{B.8}$$

and

$$\begin{aligned}
&\mathsf{Var}\left\{\mathsf{E}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{s}, \boldsymbol{x})|\boldsymbol{x}\right\} = \mathsf{Var}\left\{\mathsf{E}\left(\boldsymbol{M}_X^{-1}\frac{1}{r}\sum_{i=1}^{r}\psi_i|\boldsymbol{s}, \boldsymbol{x}\right)|\boldsymbol{x}\right\} \\
&= \mathsf{Var}\left[\boldsymbol{M}_X^{-1}\mathsf{E}\left\{\frac{(y_i - p_i(\hat{\boldsymbol{\beta}}_{MLE}))\boldsymbol{x}_i}{n\pi_i}\Big|\boldsymbol{s}, \boldsymbol{x}\right\}\Big|\boldsymbol{x}\right] \\
&= \mathsf{Var}\left\{\frac{1}{n}\boldsymbol{M}_X^{-1}\sum_{i=1}^{n}(p_i(\hat{\boldsymbol{\alpha}}_{MLE}) - p_i(\hat{\boldsymbol{\beta}}_{MLE}))\Big|\boldsymbol{x}\right\} = 0
\end{aligned} \tag{B.9}$$

Therefore, combining (B.7), (B.8) and (B.9),

$$\begin{aligned}
&\mathsf{Var}\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{x}\right) = \\
&\frac{1}{n^2 r}\boldsymbol{M}_X^{-1}\left[\sum_{i=1}^{n}\left\{p_i(\hat{\boldsymbol{\alpha}}_{MLE}) - 2 * p_i(\hat{\boldsymbol{\alpha}}_{MLE})p_i(\hat{\boldsymbol{\beta}}_{MLE}) + p_i(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}\boldsymbol{x}_i\boldsymbol{x}_i^T\left(\frac{1}{\pi_i} - 1\right)\right]\boldsymbol{M}_X^{-1} \\
&+ \frac{1}{n^2}\boldsymbol{M}_X^{-1}\sum_{i=1}^{n}p_i(\hat{\boldsymbol{\alpha}}_{MLE})(1 - p_i(\hat{\boldsymbol{\alpha}}_{MLE}))\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{M}_X^{-1}
\end{aligned} \tag{B.10}$$

Then taking the trace, we have

$$
\begin{aligned}
&\text{Trace}\left\{\text{Var}\left(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{MLE}|\boldsymbol{x}\right)\right\}= \\
&\quad\frac{1}{n^2 r}\sum_{i=1}^{n} tr\left[\left\{p_i(\hat{\boldsymbol{\alpha}}_{MLE})-2*p_i(\hat{\boldsymbol{\alpha}}_{MLE})p_i(\hat{\boldsymbol{\beta}}_{MLE})+p_i(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}\boldsymbol{M}_X^{-1}\boldsymbol{x}_i\boldsymbol{x}_i\boldsymbol{M}_X^{-1}\left(\frac{1}{\pi_i}-1\right)\right] \\
&\quad+\frac{1}{n^2}\sum_{i=1}^{n} tr\left\{p_i(\hat{\boldsymbol{\alpha}}_{MLE})(1-p_i(\hat{\boldsymbol{\alpha}}_{MLE}))\boldsymbol{M}_X^{-1}\boldsymbol{x}_i\boldsymbol{x}_i\boldsymbol{M}_X^{-1}\right\} \\
&\quad=\frac{1}{n^2 r}\sum_{i=1}^{n} tr\left[\frac{\left\{p_i(\hat{\boldsymbol{\alpha}}_{MLE})-2*p_i(\hat{\boldsymbol{\alpha}}_{MLE})p_i(\hat{\boldsymbol{\beta}}_{MLE})+p_i(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}}{\pi_i}\boldsymbol{M}_X^{-1}\boldsymbol{x}_i\boldsymbol{x}_i\boldsymbol{M}_X^{-1}\right]+c
\end{aligned}
$$

$$
\begin{aligned}
&=\frac{1}{n^2 r}\sum_{i=1}^{n}\frac{\left\{p_i(\hat{\boldsymbol{\alpha}}_{MLE})-2*p_i(\hat{\boldsymbol{\alpha}}_{MLE})p_i(\hat{\boldsymbol{\beta}}_{MLE})+p_i(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}}{\pi_i}||\boldsymbol{M}_X^{-1}\boldsymbol{x}_i||^2+c \\
&=\frac{1}{n^2 r}\sum_{i=1}^{n}\pi_i\sum_{i=1}^{n}\frac{\left\{p_i(\hat{\boldsymbol{\alpha}}_{MLE})-2*p_i(\hat{\boldsymbol{\alpha}}_{MLE})p_i(\hat{\boldsymbol{\beta}}_{MLE})+p_i(\hat{\boldsymbol{\beta}}_{MLE})^2\right\}}{\pi_i}||\boldsymbol{M}_X^{-1}\boldsymbol{x}_i||^2+c \\
&\geq\frac{1}{n^2 r}\left\{\sum_{i=1}^{n}\sqrt{p_i(\hat{\boldsymbol{\alpha}}_{MLE})-2*p_i(\hat{\boldsymbol{\alpha}}_{MLE})p_i(\hat{\boldsymbol{\beta}}_{MLE})+p_i(\hat{\boldsymbol{\beta}}_{MLE})^2}||\boldsymbol{M}_X^{-1}\boldsymbol{x}_i||\right\}^2+c \qquad \text{(B.11)}
\end{aligned}
$$

where the last Cauchy-Schwarz inequality holds if and only if

$$
\pi_i=\pi_{i,s\text{AUG}}\propto\sqrt{p_i(\hat{\boldsymbol{\alpha}}_{MLE})-2*p_i(\hat{\boldsymbol{\alpha}}_{MLE})p_i(\hat{\boldsymbol{\beta}}_{MLE})+p_i(\hat{\boldsymbol{\beta}}_{MLE})^2}||\boldsymbol{M}_X^{-1}\boldsymbol{x}_i|| \qquad \text{(B.12)}
$$

B.3. Proof of Proposition 2

To measure the extent to which $\pi_{i,s\text{SUB}}$ serves as a proxy measure for the influence of observation i on estimation of $\tilde{\boldsymbol{\beta}}$, we propose projecting the influence function for $\tilde{\gamma}$ onto the tangent space spanned by the influence function for $\tilde{\boldsymbol{\beta}}$, as follows:

Let

$$
\varphi_{\hat{\boldsymbol{\beta}}_{MLE}}=(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{MLE})=\tilde{\boldsymbol{M}}_X^{-1}\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n}+O_{P|\mathscr{F}}(r^{-1})
$$

and

$$
\varphi_{\hat{\gamma}_{MLE}}=(\tilde{\gamma}-\hat{\gamma}_{MLE})=\tilde{\boldsymbol{Q}}_X^{-1}\frac{\dot{\ell}^*(\hat{\gamma}_{MLE})}{n}+O_{P|\mathscr{F}}(r^{-1})
$$

be the influence functions for $\tilde{\boldsymbol{\beta}}$ and $\tilde{\gamma}$ respectively.

To find the projection of $\varphi_{\hat{\boldsymbol{\gamma}}_{MLE}}$ onto the space spanned by $\varphi_{\hat{\boldsymbol{\beta}}_{MLE}}$, denoted as $\Pi\big(\varphi^*(\hat{\boldsymbol{\gamma}}_{MLE})|\wedge_{\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})}\big)$, note that

$$E\left[\left\{\varphi^*(\boldsymbol{\gamma})-\Pi\big(\varphi^*(\hat{\boldsymbol{\gamma}}_{MLE})|\wedge_{\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})}\big)\right\}\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})\Big|\boldsymbol{x}_i^*\right]=0 \tag{B.13}$$

Let $\Pi\big(\varphi^*(\hat{\boldsymbol{\gamma}}_{MLE})|\wedge_{\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})}\big)=V\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})$. Then

$$\begin{aligned}
&E\left[\left\{\varphi^*(\hat{\boldsymbol{\gamma}}_{MLE})-V\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})\right\}\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})^T|\boldsymbol{x}_i^*\right]=0\\
&\Rightarrow\ E\left\{\varphi^*(\hat{\boldsymbol{\gamma}}_{MLE})\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})^T|\boldsymbol{x}_i^*\right\}=VE\left\{\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})^T|\boldsymbol{x}_i^*\right\}\\
&\Rightarrow\ V=E\left\{\varphi^*(\hat{\boldsymbol{\gamma}}_{MLE})\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})^T|\boldsymbol{x}_i^*\right\}E\left\{\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})^T|\boldsymbol{x}_i^*\right\}^{-1}\\
&\quad=E\left\{\tilde{\boldsymbol{Q}}_X^{-1}\frac{\dot{\ell}^*(\hat{\boldsymbol{\gamma}}_{MLE})}{n}\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n}\tilde{\boldsymbol{M}}_X^{-1}\right\}E\left\{\tilde{\boldsymbol{M}}_X^{-1}\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n}\frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n}\tilde{\boldsymbol{M}}_X^{-1}\right\}^{-1}\\
&\quad=E\left[\tilde{\boldsymbol{Q}}_X^{-1}\left\{\frac{1}{r^2n^2}\sum_{i=1}^r\frac{(s_i^*-p_i^*(\boldsymbol{\gamma}))\boldsymbol{x}_i^*}{\pi_i^*}\sum_{i=1}^r\frac{(y_i^*-p_i^*(\boldsymbol{\beta}))\boldsymbol{x}_i^*}{\pi_i^*}\right\}\tilde{\boldsymbol{M}}_X^{-1}\right]\\
&\qquad\qquad E\left[\tilde{\boldsymbol{M}}_X^{-1}\left\{\frac{1}{r^2n^2}\sum_{i=1}^r\frac{(y_i^*-p_i^*(\boldsymbol{\beta}))\boldsymbol{x}_i^*}{\pi_i^*}\sum_{i=1}^r\frac{(y_i^*-p_i^*(\boldsymbol{\beta}))\boldsymbol{x}_i^*}{\pi_i^*}\right\}\tilde{\boldsymbol{M}}_X^{-1}\right]^{-1}\\
&\quad=E\left[\tilde{\boldsymbol{Q}}_X^{-1}\left\{\frac{1}{r^2n^2}\sum_{i=1}^r\frac{(s_i^*-p_i^*(\boldsymbol{\gamma}))(y_i^*-p_i^*(\boldsymbol{\beta}))\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}}{\pi_i^*}\right\}\tilde{\boldsymbol{M}}_X^{-1}\right]\\
&\qquad\qquad E\left[\tilde{\boldsymbol{M}}_X^{-1}\left\{\frac{1}{r^2n^2}\sum_{i=1}^r\frac{(y_i^*-p_i^*(\boldsymbol{\beta}))^2\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}}{\pi_i^*}\right\}\tilde{\boldsymbol{M}}_X^{-1}\right]^{-1}\\
&\quad\text{(since cov}(s_i^*,y_i^*)=\text{cov}(y_i^*,y_j^*)=0\text{ for }i\neq j)\\
&\quad=E\left[\tilde{\boldsymbol{Q}}_X^{-1}\left\{\frac{1}{rn}\sum_{i=1}^r\frac{(s_i^*-p_i^*(\boldsymbol{\gamma}))(y_i^*-p_i^*(\boldsymbol{\beta}))\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}}{\pi_i^*}\right\}\tilde{\boldsymbol{M}}_X^{-1}\right]\\
&\qquad\qquad E\left[\tilde{\boldsymbol{M}}_X^{-1}\left\{\frac{1}{rn}\sum_{i=1}^r\frac{(y_i^*-p_i^*(\boldsymbol{\beta}))^2\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}}{\pi_i^*}\right\}\tilde{\boldsymbol{M}}_X^{-1}\right]^{-1}\\
&\quad=\left(\tilde{\boldsymbol{Q}}_X^{-1}\left[\frac{1}{rn}\sum_{i=1}^r\frac{\{E(y_i^*s_i^*|\boldsymbol{x}_i^*)-p_i^*(\boldsymbol{\beta})E(s_i^*|\boldsymbol{x}_i^*)-p_i^*(\boldsymbol{\gamma})E(y_i^*|\boldsymbol{x}_i^*)+p_i^*(\boldsymbol{\beta})p_i^*(\boldsymbol{\gamma})\}\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}}{\pi_i^*}\right]\tilde{\boldsymbol{M}}_X^{-1}\right)\\
&\qquad\qquad\left(\tilde{\boldsymbol{M}}_X^{-1}\left[\frac{1}{rn}\sum_{i=1}^r\frac{\{E(y_i^{*2}|\boldsymbol{x}_i^*)-2p_i^*(\boldsymbol{\beta})E(y_i^*|\boldsymbol{x}_i^*)+(p_i^*(\boldsymbol{\beta}))^2\}\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}}{\pi_i^*}\right]\tilde{\boldsymbol{M}}_X^{-1}\right)^{-1}\\
&\quad=\left(\tilde{\boldsymbol{Q}}_X^{-1}\left[\frac{1}{rn}\sum_{i=1}^r\frac{\{E(y_i^*s_i^*|\boldsymbol{x}_i^*)-p_i^*(\boldsymbol{\beta})p_i^*(\boldsymbol{\gamma})\}\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}}{\pi_i^*}\right]\tilde{\boldsymbol{M}}_X^{-1}\right)\\
&\qquad\qquad\left(\tilde{\boldsymbol{M}}_X^{-1}\left[\frac{1}{rn}\sum_{i=1}^r\frac{\{E(y_i^{*2}|\boldsymbol{x}_i^*)-(p_i^*(\boldsymbol{\beta}))^2\}\boldsymbol{x}_i^*\boldsymbol{x}_i^{*T}}{\pi_i^*}\right]\tilde{\boldsymbol{M}}_X^{-1}\right)^{-1}
\end{aligned}$$

$$
\begin{aligned}
&= \left(\tilde{\boldsymbol{Q}}_X^{-1} \left[\frac{1}{rn} \sum_{i=1}^{r} \frac{\{E(y_i^* s_i^* | \boldsymbol{x}_i^*) - p_i^*(\boldsymbol{\beta}) p_i^*(\boldsymbol{\gamma})\} \boldsymbol{x}_i^* \boldsymbol{x}_i^{*T}}{\pi_i^*} \right] \tilde{\boldsymbol{M}}_X^{-1} \right) \\
&\qquad \left(\tilde{\boldsymbol{M}}_X^{-1} \left[\frac{1}{rn} \sum_{i=1}^{r} \frac{\{p_i^*(\boldsymbol{\beta})(1 - p_i^*(\boldsymbol{\beta}))\} \boldsymbol{x}_i^* \boldsymbol{x}_i^{*T}}{\pi_i^*} \right] \tilde{\boldsymbol{M}}_X^{-1} \right)^{-1} \\
&= \left(\tilde{\boldsymbol{Q}}_X^{-1} \left[\frac{1}{rn} \sum_{i=1}^{r} \frac{\{E(y_i^* s_i^* | \boldsymbol{x}_i^*) - p_i^*(\boldsymbol{\beta}) p_i^*(\boldsymbol{\gamma})\} \boldsymbol{x}_i^* \boldsymbol{x}_i^{*T}}{\pi_i^*} \right] \tilde{\boldsymbol{M}}_X^{-1} \right) \left(\tilde{\boldsymbol{M}}_X^{-1} \tilde{\boldsymbol{M}}_X \tilde{\boldsymbol{M}}_X^{-1} \right)^{-1} \\
&= \tilde{\boldsymbol{Q}}_X^{-1} \left[\frac{1}{rn} \sum_{i=1}^{r} \frac{\{E(y_i^* s_i^* | \boldsymbol{x}_i^*) - p_i^*(\boldsymbol{\beta}) p_i^*(\boldsymbol{\gamma})\} \boldsymbol{x}_i^* \boldsymbol{x}_i^{*T}}{\pi_i^*} \right] \tilde{\boldsymbol{M}}_X^{-1} \tilde{\boldsymbol{M}}_X \\
&= \tilde{\boldsymbol{Q}}_X^{-1} \left[\frac{1}{rn} \sum_{i=1}^{r} \frac{\{E(y_i^* s_i^* | \boldsymbol{x}_i^*) - p_i^*(\boldsymbol{\beta}) p_i^*(\boldsymbol{\gamma})\} \boldsymbol{x}_i^* \boldsymbol{x}_i^{*T}}{\pi_i^*} \right]
\end{aligned}
$$

We can select a working model for $E[y_i^* s_i^* | \boldsymbol{x}_i^*]$, for example

$$
\text{logit}\{Pr(s_i = 1, y_i = 1 | \boldsymbol{x}_i)\} = \boldsymbol{x}_i^T \boldsymbol{\eta} \tag{B.14}
$$

Then

$$
\begin{aligned}
\Pi\big(\varphi^*(\hat{\boldsymbol{\gamma}}_{MLE}) | \wedge_{\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})}\big) &= \left(\tilde{\boldsymbol{Q}}_X^{-1} \left[\frac{1}{rn} \sum_{i=1}^{r} \frac{\{p_i^*(\boldsymbol{\eta}) - p_i^*(\boldsymbol{\beta}) p_i^*(\boldsymbol{\gamma})\} \boldsymbol{x}_i^* \boldsymbol{x}_i^{*T}}{\pi_i^*} \right] \right) \varphi^*(\hat{\boldsymbol{\beta}}_{MLE}) \\
&= \left(\tilde{\boldsymbol{Q}}_X^{-1} \left[\frac{1}{rn} \sum_{i=1}^{r} \frac{\{p_i^*(\boldsymbol{\eta}) - p_i^*(\boldsymbol{\beta}) p_i^*(\boldsymbol{\gamma})\} \boldsymbol{x}_i^* \boldsymbol{x}_i^{*T}}{\pi_i^*} \right] \right) \tilde{\boldsymbol{M}}_X^{-1} \frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n} \\
&= \tilde{\boldsymbol{Q}}_X^{-1} \tilde{\boldsymbol{A}}_X \tilde{\boldsymbol{M}}_X^{-1} \frac{\dot{\ell}^*(\hat{\boldsymbol{\beta}}_{MLE})}{n}
\end{aligned} \tag{B.15}
$$

where $\tilde{\boldsymbol{A}}_X = \dfrac{1}{rn} \displaystyle\sum_{i=1}^{r} \frac{\{p_i^*(\boldsymbol{\eta}) - p_i^*(\boldsymbol{\beta}) p_i^*(\boldsymbol{\gamma})\} \boldsymbol{x}_i^* \boldsymbol{x}_i^{*T}}{\pi_i^*}$

Note that this projected influence function effectively scales the $\varphi^*(\hat{\boldsymbol{\beta}}_{MLE})$ by $\tilde{\boldsymbol{Q}}_X^{-1} \tilde{\boldsymbol{A}}_X$, giving a geometric representation of the component of $\varphi_{\hat{\boldsymbol{\beta}}_{MLE}}$ that runs along the direction of $\varphi_{\hat{\boldsymbol{\gamma}}_{MLE}}$.

B.4. Derivation of Surrogate Substitution Weights

Let us consider $\boldsymbol{\gamma}$, the log odds ratio in the model for logit$(Pr(s_i | \boldsymbol{x}_i)$. $\tilde{\boldsymbol{\gamma}})$ is the solution to the weighted score equation

$$
\dot{\ell}^*(\tilde{\boldsymbol{\gamma}}) = \frac{1}{r} \sum_{i=1}^{r} \frac{\{s_i^* - p_i^*(\tilde{\boldsymbol{\gamma}})\} \boldsymbol{x}_i^*}{\pi_i^*} = 0
$$

Similar to how we derived the asymptotic distribution for $\tilde{\boldsymbol{\beta}}$ in Section B.1, it can be shown that

$$
r^{1/2} \boldsymbol{V}_{\gamma}^{-1/2} (\tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}}_{MLE}) \xrightarrow{D|\mathscr{F}} N(0, \boldsymbol{I}) \tag{B.16}
$$

where $\boldsymbol{V}_\gamma = \boldsymbol{Q}_X^{-1}\text{Var}[\psi_{\gamma,i}|\mathscr{F}]\boldsymbol{Q}_X^{-1}$

$$\psi_{\gamma,i} = \frac{\{s_i - p_i(\hat{\gamma}_{MLE})\}\,\boldsymbol{x}_i}{n\pi_i}$$

$$\boldsymbol{Q}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} p_i(\hat{\gamma}_{MLE})\,\{1 - p_i(\hat{\gamma}_{MLE})\}\,\boldsymbol{x}_i\boldsymbol{x}_i^T$$

Then we can derive the optimal weights for $\tilde{\gamma}$ by taking the trace of $\boldsymbol{V}_\gamma$

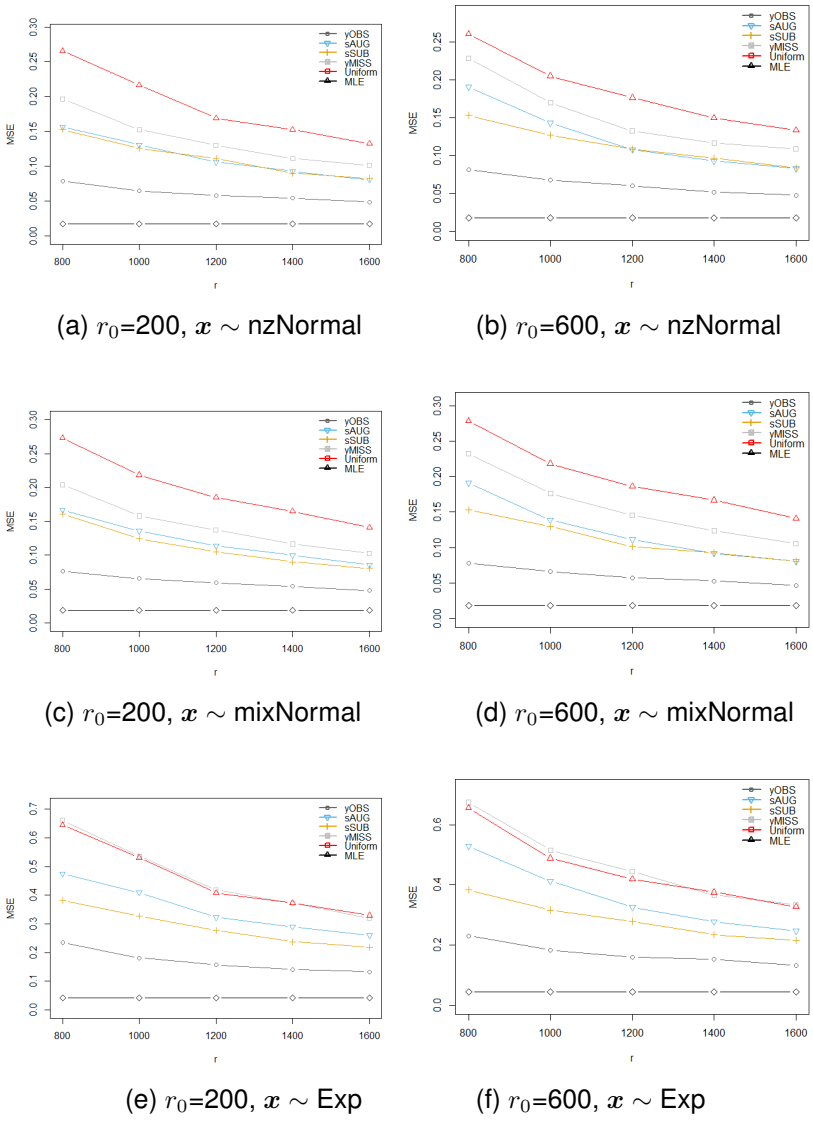
$$\begin{aligned}
\text{Trace}(\boldsymbol{V}_\gamma) &= \frac{1}{n^2 r}\sum_{i=1}^{n} tr\left[\frac{\{s_i - p_i(\hat{\gamma}_{MLE})\}^2}{\pi_i}\boldsymbol{Q}_X^{-1}\boldsymbol{x}_i\boldsymbol{x}_i\boldsymbol{Q}_X^{-1}\right] \\
&= \frac{1}{n^2 r}\sum_{i=1}^{n}\frac{\{s_i - p_i(\hat{\gamma}_{MLE})\}^2}{\pi_i}||\boldsymbol{Q}_X^{-1}\boldsymbol{x}_i||^2 \\
&= \frac{1}{n^2 r}\sum_{i=1}^{n}\pi_i\sum_{i=1}^{n}\frac{\{s_i - p_i(\hat{\gamma}_{MLE})\}^2}{\pi_i}||\boldsymbol{Q}_X^{-1}\boldsymbol{x}_i||^2 \\
&\geq \frac{1}{n^2 r}\left\{\sum_{i=1}^{n}|s_i - p_i(\hat{\gamma}_{MLE})|||\boldsymbol{Q}_X^{-1}\boldsymbol{x}_i||\right\}^2
\end{aligned} \tag{B.17}$$

where the last Cauchy-Schwarz inequality holds if and only if

$$\pi_i = \pi_{i,s\text{SUB}} \propto |s_i - p_i(\hat{\gamma}_{MLE})|||\boldsymbol{Q}_X^{-1}\boldsymbol{x}_i|| \tag{B.18}$$

B.5. Additional Simulation results

Figure B.1: Empirical mean squared error of $\tilde{\beta}$ using different sampling weights, over 500 replicates. $(se, sp)_{x_1 \leq 0.4} = (0.95, 0.90)$, $(se, sp)_{x_1 > 0.4} = (0.85, 0.80)$, $n = 10,000$

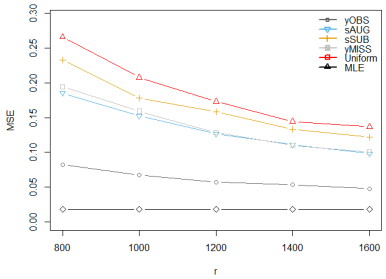


(a) r_0=200, $\boldsymbol{x} \sim$ nzNormal (b) r_0=600, $\boldsymbol{x} \sim$ nzNormal

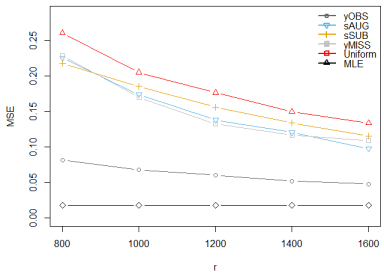
(c) r_0=200, $\boldsymbol{x} \sim$ mixNormal (d) r_0=600, $\boldsymbol{x} \sim$ mixNormal

(e) r_0=200, $\boldsymbol{x} \sim$ Exp (f) r_0=600, $\boldsymbol{x} \sim$ Exp

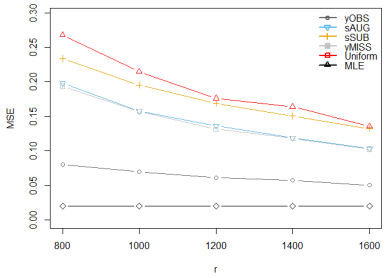
Figure B.2: Empirical mean squared error of $\tilde{\beta}$ using different sampling weights, over 500 replicates. $(se, sp)_{x_1 \leq 0.4} = (0.70, 0.65)$, $(se, sp)_{x_1 > 0.4} = (0.60, 0.55)$, $n = 10,000$



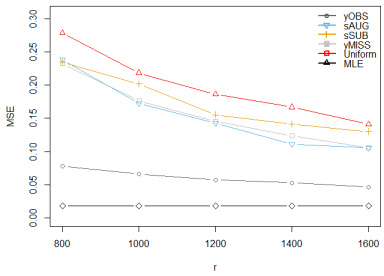
(a) r_0=200, $\boldsymbol{x} \sim$ nzNormal



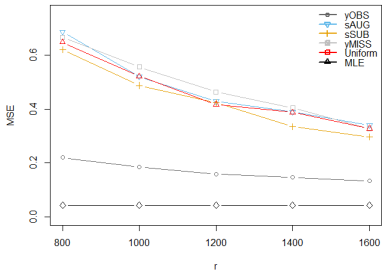
(b) r_0=600, $\boldsymbol{x} \sim$ nzNormal



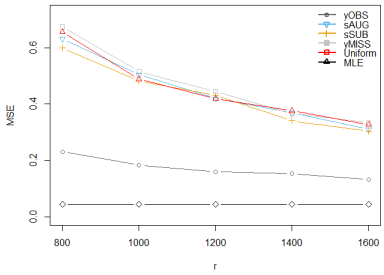
(c) r_0=200, $\boldsymbol{x} \sim$ mixNormal



(d) r_0=600, $\boldsymbol{x} \sim$ mixNormal



(e) r_0=200, $\boldsymbol{x} \sim$ Exp



(f) r_0=600, $\boldsymbol{x} \sim$ Exp

BIBLIOGRAPHY

Abul-Husn, NS and Kenny, EE (2019). Personalized medicine and the power of electronic health records. *Cell* 177.1, 58–69.

Al-Durra, M, Nolan, RP, Seto, E, Cafazzo, JA, and Eysenbach, G (2018). Nonpublication Rates and Characteristics of Registered Randomized Clinical Trials in Digital Health: Cross-Sectional Analysis. *Journal of medical Internet research* 20.12, e11924.

Alper, BS, Hand, JA, Elliott, SG, Kinkade, S, Hauan, MJ, Onion, DK, and Sklar, BM (2004). How much effort is needed to keep up with the literature relevant for primary care? *Journal of the Medical Library association* 92.4, 429.

Bardy, AH (1998). Bias in reporting clinical trials. *British journal of clinical pharmacology* 46.2, 147–150.

Beesley, LJ and Mukherjee, B (2020). Statistical inference for association studies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*.

Begg, CB and Mazumdar, M (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088–1101.

Berlin, JA and Golub, RM (2014). Meta-analysis as evidence: building a better pyramid. *Jama* 312.6, 603–606.

Boudreau, DM, Yu, O, Chubak, J, Wirtz, HS, Bowles, EJA, Fujii, M, and Buist, DS (2014). Comparative safety of cardiovascular medication use and breast cancer outcomes among women with early stage breast cancer. *Breast cancer research and treatment* 144.2, 405–416.

Bower, JK, Patel, S, Rudy, JE, and Felix, AS (2017). Addressing bias in electronic health record-based surveillance of cardiovascular disease risk: finding the signal through the noise. *Current epidemiology reports* 4.4, 346–352.

Carpenter, JR, Schwarzer, G, Rücker, G, and Künstler, R (2009). Empirical evaluation showed that the Copas selection model provided a useful summary in 80% of meta-analyses. *Journal of clinical epidemiology* 62.6, 624–631.

Cartabellotta, A and Tilson, JK (2019). The ecosystem of evidence cannot thrive without efficiency of knowledge generation, synthesis, and translation. *Journal of clinical epidemiology* 110, 90–95.

Chakrabortty, A, Cai, T, et al. (2018). Efficient and adaptive linear regression in semi-supervised settings. *Annals of Statistics* 46.4, 1541–1572.

Chan, N (1982). A-optimality for regression designs. *Journal of Mathematical Analysis and Applications* 87.1, 45–50.

Chen, Y, Hong, C, and Riley, RD (2015). An alternative pseudolikelihood method for multivariate random-effects meta-analysis. *Statistics in medicine* 34.3, 361–380.

Chen, Y, Wang, J, Chubak, J, and Hubbard, RA (2019). Inflation of type I error rates due to differential misclassification in EHR-derived outcomes: Empirical illustration using breast cancer recurrence. *Pharmacoepidemiology and drug safety* 28.2, 264–268.

Cheng, D, Ananthakrishnan, AN, and Cai, T (2020). Robust and efficient semi-supervised estimation of average treatment effects with application to electronic health records data. *Biometrics*, 413–423.

Chootrakool, H, Shi, JQ, and Yue, R (2011). Meta-analysis and sensitivity analysis for multi-arm trials with selection bias. *Statistics in medicine* 30.11, 1183–1198.

Chubak, J, Yu, O, Pocobelli, G, Lamerato, L, Webster, J, Prout, MN, Ulcickas Yood, M, Barlow, WE, and Buist, DS (2012). Administrative data algorithms to identify second breast cancer events following early-stage invasive breast cancer. *Journal of the National Cancer Institute* 104.12, 931–940.

Chubak, J, Yu, O, Ziebell, RA, Bowles, EJA, Sterrett, AT, Fujii, MM, Boggs, JM, Burnett-Hartman, AN, Boudreau, DM, Chen, L, et al. (2018). Risk of colon cancer recurrence in relation to diabetes. *Cancer Causes & Control* 29.11, 1093–1103.

Classen, DC, Pestotnik, SL, Evans, RS, and Burke, JP (1991). Computerized surveillance of adverse drug events in hospital patients. *Jama* 266.20, 2847–2851.

Cohn, LD and Becker, BJ (2003). How meta-analysis increases statistical power. *Psychological methods* 8.3, 243.

Cooper, H, DeNeve, K, and Charlton, K (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods* 2.4, 447.

Copas, J and Shi, JQ (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical methods in medical research* 10.4, 251–265.

Copas, J and Shi, JQ (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics* 1.3, 247–262.

De Angelis, C, Drazen, JM, Frizelle, FA, Haug, C, Hoey, J, Horton, R, Kotzin, S, Laine, C, Marusic, A, Overbeke, AJP, et al. (2005). Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *Circulation* 111.10, 1337–1338.

Denaxas, S, Direk, K, Gonzalez-Izquierdo, A, Pikoula, M, Cakiroglu, A, Moore, J, Hemingway, H, and Smeeth, L (2017). Methods for enhancing the reproducibility of biomedical research findings using electronic health records. *BioData mining* 10.1, 31.

Deville, JC, Särndal, CE, and Sautory, O (1993). Generalized raking procedures in survey sampling. *Journal of the American statistical Association* 88.423, 1013–1020.

Dickersin, K and Min, Y (1993). *NIH clinical trials and publication bias.*

Drineas, P, Mahoney, MW, and Muthukrishnan, S (2006). "Sampling algorithms for l 2 regression and applications". In: *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, 1127–1136.

Duan, R, Cao, M, Wu, Y, Huang, J, Denny, JC, Xu, H, and Chen, Y (2016). “An empirical study for impacts of measurement errors on EHR based association studies”. In: *AMIA Annual Symposium Proceedings*. Vol. 2016. American Medical Informatics Association, 1764.

Duval, S and Tweedie, R (2000). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association* 95.449, 89–98.

Egger, M, Smith, GD, Schneider, M, and Minder, C (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj* 315.7109, 629–634.

Embi, PJ and Payne, PR (2013). Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. *Medical care* 51, S87–S91.

Ferryman, K and Pitcan, M (2018). Fairness in precision medicine. *Data & Society* 1.

Firth, D (1993). Bias reduction of maximum likelihood estimates. *Biometrika* 80.1, 27–38.

Fithian, W and Hastie, T (2014). Local case-control sampling: Efficient subsampling in imbalanced data sets. *Annals of statistics* 42.5, 1693.

Forrest, CB, Margolis, PA, Bailey, LC, Marsolo, K, Del Beccaro, MA, Finkelstein, JA, Milov, DE, Vieland, VJ, Wolf, BA, Yu, FB, et al. (2014). PEDSnet: a national pediatric learning health system. *Journal of the American Medical Informatics Association* 21.4, 602–606.

Fry, A, Littlejohns, TJ, Sudlow, C, Doherty, N, Adamska, L, Sprosen, T, Collins, R, and Allen, NE (2017). Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American journal of epidemiology* 186.9, 1026–1034.

Galbraith, R (1988). Graphical display of estimates having differing standard errors. *Technometrics* 30.3, 271–281.

Gianfrancesco, MA, Tamang, S, Yazdany, J, and Schmajuk, G (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine* 178.11, 1544–1547.

Greenhalgh, T, Howick, J, and Maskrey, N (2014). Evidence based medicine: a movement in crisis? *Bmj* 348.

Guyatt, G, Cairns, J, Churchill, D, Cook, D, Haynes, B, Hirsh, J, Irvine, J, Levine, M, Levine, M, Nishikawa, J, et al. (1992). Evidence-based medicine: a new approach to teaching the practice of medicine. *Jama* 268.17, 2420–2425.

Hamamura, FD, Withy, K, and Hughes, K (2017). Identifying barriers in the use of electronic health Records in Hawai ‘i. *Hawai’i Journal of Medicine & Public Health* 76.3 Suppl 1, 28.

Haneuse, S and Daniels, M (2016). A general framework for considering selection bias in EHR-based studies: what data are observed and why? *eGEMs* 4.1.

Hassett, MJ, Uno, H, Cronin, AM, Carroll, NM, Hornbrook, MC, and Ritzwoller, D (2017). Detecting lung and colorectal cancer recurrence using structured clinical/administrative data to enable outcomes research and population health management. *Medical care* 55.12, e88.

Hedges, LV (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics* 9.1, 61–85.

Hedges, LV (1992). Modeling publication selection effects in meta-analysis. *Statistical Science*, 246–255.

Hing, E and Burt, CW (2009). Are there patient disparities when electronic health records are adopted? *Journal of health care for the poor and underserved* 20.2, 473–488.

Hripcsak, G and Albers, DJ (2012). Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* 20.1, 117–121.

Hubbard, RA, Huang, J, Harton, J, Oganisian, A, Choi, G, Utidjian, L, Eneli, I, Bailey, LC, and Chen, Y (2019). A Bayesian latent class approach for EHR-based phenotyping. *Statistics in medicine* 38.1, 74–87.

Hubbard, RA, Tong, J, Duan, R, and Chen, Y (2020). Reducing Bias Due to Outcome Misclassification for Epidemiologic Studies Using EHR-derived Probabilistic Phenotypes. *Epidemiology* 31.4, 542–550.

Ibrahim, SA, Charlson, ME, and Neill, DB (2020). Big Data Analytics and the Struggle for Equity in Health Care: The Promise and Perils. *Health Equity* 4.1, 99–101.

Iyengar, S and Greenhouse, JB (1988). Selection models and the file drawer problem. *Statistical Science*, 109–117.

Jensen, PB, Jensen, LJ, and Brunak, S (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* 13.6, 395.

Jin, ZC, Zhou, XH, and He, J (2015). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in medicine* 34.2, 343–360.

Johnson, VE, Payne, RD, Wang, T, Asher, A, and Mandal, S (2017). On the reproducibility of psychological science. *Journal of the American Statistical Association* 112.517, 1–10.

Jones, CW, Handler, L, Crowell, KE, Keil, LG, Weaver, MA, and Platts-Mills, TF (2013). Non-publication of large randomized clinical trials: cross sectional analysis. *Bmj* 347, f6104.

Kaplan, RM, Chambers, DA, and Glasgow, RE (2014). Big data and large sample size: a cautionary note on the potential for bias. *Clinical and translational science* 7.4, 342–346.

Karlis, D and Xekalaki, E (2003). Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics & Data Analysis* 41.3-4, 577–590.

Kemper, AR, Uren, RL, and Clark, SJ (2006). Adoption of electronic health records in primary care pediatric practices. *Pediatrics* 118.1, e20–e24.

Khoo, AL, Zhou, HJ, Teng, M, Lin, L, Zhao, YJ, Soh, LB, Mok, YM, Lim, BP, and Gwee, KP (2015). Network meta-analysis and cost-effectiveness analysis of new generation antidepressants. *CNS drugs* 29.8, 695–712.

Kicinski, M, Springate, DA, and Kontopantelis, E (2015). Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Statistics in medicine* 34.20, 2781–2793.

Kilbridge, PM, Campbell, UC, Cozart, HB, and Mojarrad, MG (2006). Automated surveillance for adverse drug events at a community hospital and an academic medical center. *Journal of the American Medical Informatics Association* 13.4, 372–377.

Kirby, JC, Speltz, P, Rasmussen, LV, Basford, M, Gottesman, O, Peissig, PL, Pacheco, JA, Tromp, G, Pathak, J, Carrell, DS, et al. (2016). PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *Journal of the American Medical Informatics Association* 23.6, 1046–1052.

Laird, N (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* 73.364, 805–811.

Lane, DM and Dunlap, WP (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology* 31.2, 107–112.

Lau, J, Ioannidis, JP, Terrin, N, Schmid, CH, and Olkin, I (2006). Evidence based medicine: The case of the misleading funnel plot. *BMJ: British Medical Journal* 333.7568, 597.

Lee, SJC, Grobe, JE, and Tiro, JA (2016). Assessing race and ethnicity data quality across cancer registries and EMRs in two hospitals. *Journal of the American Medical Informatics Association* 23.3, 627–634.

Light, RJ and Pillemer, DB (1984). *Summing up*. Harvard University Press.

López, MM, Bevans, M, Wehrlen, L, Yang, L, and Wallen, G (2017). Discrepancies in race and ethnicity documentation: a potential barrier in identifying racial and ethnic disparities. *Journal of racial and ethnic health disparities* 4.5, 812–818.

Louis, TA (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 226–233.

Lu, G and Ades, A (2004). Combination of direct and indirect evidence in mixed treatment comparisons. *Statistics in medicine* 23.20, 3105–3124.

Lumley, T (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in medicine* 21.16, 2313–2324.

Ma, P, Mahoney, MW, and Yu, B (2015). A statistical perspective on algorithmic leveraging. *The Journal of Machine Learning Research* 16.1, 861–911.

Ma, P and Sun, X (2015). Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics* 7.1, 70–76.

Mack, D, Zhang, S, Douglas, M, Sow, C, Strothers, H, and Rust, G (2016). Disparities in primary care EHR adoption rates. *Journal of health care for the poor and underserved* 27.1, 327.

Mahoney, MW and Drineas, P (2009). CUR matrix decompositions for improved data analysis. *Proceedings of the National Academy of Sciences* 106.3, 697–702.

Makam, AN and Nguyen, OK (2017). An evidence-based medicine approach to antihyperglycemic therapy in diabetes mellitus to overcome overtreatment. *Circulation* 135.2, 180–195.

Manaktala, S and Claypool, SR (2017). Evaluating the impact of a computerized surveillance algorithm and decision support system on sepsis mortality. *Journal of the American Medical Informatics Association* 24.1, 88–95.

Marks-Anglin, A and Chen, Y (2020a). A Historical Review of Publication Bias. *Research Synthesis Methods*. DOI: 10.1002/jrsm.1452.

Marks-Anglin, A and Chen, Y (2020b). Small-study effects: current practice and challenges for future research. *Statistics and its Interface* 13.4, 475–484.

Marks-Anglin, AK, Luo, C, Hubbard, R, and Chen, Y (2021). Surrogate-assisted sampling for cost-efficient validation of electronic health record outcomes. *arXiv preprint*.

Martin, S, Wagner, J, Lupulescu-Mann, N, Ramsey, K, Cohen, AA, Graven, P, Weiskopf, NG, and Dorr, DA (2017). Comparison of EHR-based diagnosis documentation locations to a gold standard for risk stratification in patients with multiple chronic conditions. *Applied clinical informatics* 8.03, 794–809.

Mavridis, D, Sutton, A, Cipriani, A, and Salanti, G (2013). A fully Bayesian application of the Copas selection model for publication bias extended to network meta-analysis. *Statistics in medicine* 32.1, 51–66.

Mavridis, D, Welton, NJ, Sutton, A, and Salanti, G (2014). A selection model for accounting for publication bias in a full network meta-analysis. *Statistics in medicine* 33.30, 5399–5412.

Menendez, ME, Janssen, SJ, and Ring, D (2016). Electronic health record-based triggers to detect adverse events after outpatient orthopaedic surgery. *BMJ Qual Saf* 25.1, 25–30.

Neuhaus, JM (1999). Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 86.4, 843–855.

Ning, J, Chen, Y, and Piao, J (2017). Maximum likelihood estimation and EM algorithm of Copas-like selection model for publication bias correction. *Biostatistics (Oxford, England)*.

Papageorgiou, SN, Tsiranidou, E, Antonoglou, GN, Deschner, J, and Jäger, A (2015). Choice of effect measure for meta-analyses of dichotomous outcomes influenced the identified heterogeneity and direction of small-study effects. *Journal of Clinical Epidemiology* 68.5, 534–541.

Parke, WR (1986). Pseudo maximum likelihood estimation: The asymptotic distribution. *The Annals of Statistics*, 355–357.

Pope, C (2003). Resisting evidence: the study of evidence-based medicine as a contemporary social movement. *Health:* 7.3, 267–282.

Rader, KA, Lipsitz, SR, Fitzmaurice, GM, Harrington, DP, Parzen, M, and Sinha, D (2017). Bias-corrected estimates for logistic regression models for complex surveys with application to the United States' Nationwide Inpatient Sample. *Statistical methods in medical research* 26.5, 2257–2269.

Rathod, KS and Wragg, A (2018). *Do patient-reported outcome measures speak for all patient subgroups: is everyone included?*

Richesson, RL, Hammond, WE, Nahm, M, Wixted, D, Simon, GE, Robinson, JG, Bauck, AE, Cifelli, D, Smerek, MM, Dickerson, J, et al. (2013). Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *Journal of the American Medical Informatics Association* 20.e2, e226–e231.

Rudin, R, Fischer, S, Shi, Y, Shekelle, P, Amill-Rosario, A, Ridgely, M, Scanlon, D, And, and Damberg, C (Jan. 2019). Trends in the Use of Clinical Decision Support by Health System-Affiliated Ambulatory Clinics in the United States, 2014-2016. *American Journal of Managed Care* 12, 4–10.

Sackett, DL, Rosenberg, WM, Gray, JM, Haynes, RB, and Richardson, WS (1996). *Evidence based medicine: what it is and what it isn't*.

Salanti, G, Higgins, JP, Ades, A, and Ioannidis, JP (2008). Evaluation of networks of randomized trials. *Statistical methods in medical research* 17.3, 279–301.

Schwarzer, G, Antes, G, and Schumacher, M (2007). A test for publication bias in meta-analysis with sparse binary data. *Statistics in medicine* 26.4, 721–733.

Schwarzer, G, Carpenter, J, and Rücker, G (2010). Empirical evaluation suggests Copas selection model preferable to trim-and-fill method for selection bias in meta-analysis. *Journal of clinical epidemiology* 63.3, 282–288.

Sim, I, Gorman, P, Greenes, RA, Haynes, RB, Kaplan, B, Lehmann, H, and Tang, PC (2001). Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association* 8.6, 527–534.

Song, F, Altman, DG, Glenny, AM, and Deeks, JJ (2003). Validity of indirect comparison for estimating efficacy of competing interventions: empirical evidence from published meta-analyses. *Bmj* 326.7387, 472.

Sterne, JA, Sutton, AJ, Ioannidis, JP, Terrin, N, Jones, DR, Lau, J, Carpenter, J, Rücker, G, Harbord, RM, Schmid, CH, et al. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *Bmj* 343, d4002.

Terrin, N, Schmid, CH, and Lau, J (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of clinical epidemiology* 58.9, 894–901.

Thorlund, K, Thabane, L, and Mills, EJ (Dec. 2013). Modelling heterogeneity variances in multiple treatment comparison meta-analysis – Are informative priors the better solution? en. *BMC Med-*

ical Research Methodology 13.1, 2. ISSN: 1471-2288. DOI: 10.1186/1471-2288-13-2. URL: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/1471-2288-13-2 (visited on 07/08/2020).

Thornton, A and Lee, P (2000). Publication bias in meta-analysis: its causes and consequences. *Journal of clinical epidemiology* 53.2, 207–216.

Tikkanen, RS, Woolhandler, S, Himmelstein, DU, Kressin, NR, Hanchate, A, Lin, MY, McCormick, D, and Lasser, KE (2017). Hospital payer and racial/ethnic mix at private academic medical centers in Boston and New York City. *International Journal of Health Services* 47.3, 460–476.

Tong, J, Huang, J, Chubak, J, Wang, X, Moore, JH, Hubbard, RA, and Chen, Y (2020). An augmented estimation procedure for EHR-based association studies accounting for differential misclassification. *Journal of the American Medical Informatics Association* 27.2, 244–253.

Trinquart, L, Dunn, AG, and Bourgeois, FT (2018). Registration of published randomized trials: a systematic review and meta-analysis. *BMC medicine* 16.1, 173.

Vandvik, PO and Brandt, L (2020). Future of evidence ecosystem series: evidence ecosystems and learning health systems: why bother? *Journal of clinical epidemiology* 123, 166–170.

Wang, H (2019). More efficient estimation for logistic regression with optimal subsamples. *Journal of machine learning research* 20.

Wang, H, Zhu, R, and Ma, P (2018). Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association* 113.522, 829–844.

Wang, L, Olson, JE, Bielinski, SJ, St Sauver, JL, Fu, S, He, H, Cicek, MS, Hathcock, MA, Cerhan, JR, and Liu, H (2020). Impact of diverse data sources on computational phenotyping. *Frontiers in Genetics* 11, 556.

Wu, Y, Warner, JL, Wang, L, Jiang, M, Xu, J, Chen, Q, Nian, H, Dai, Q, Du, X, Yang, P, et al. (2019). 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, 1–9.

Zhang, G, Beesley, LJ, Mukherjee, B, and Shi, X (2020). Patient Recruitment Using Electronic Health Records Under Selection Bias: a Two-phase Sampling Framework. *arXiv:2011.06663*.

Zhang, T, Ning, Y, and Ruppert, D (2019). Optimal Sampling for Generalized Linear Models under Measurement Constraints. *arXiv preprint arXiv:1907.07309*.