NEW STATISTICAL METHODS FOR THE EVALUATION OF EFFECTIVENESS AND SAFETY OF A MEDICAL INTERVENTION IN USING OBSERVATIONAL DATA

Jia Zhan

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Biostatistics,

Indiana University

March 2017

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Changyu Shen, PhD, Co-Chair

Xiaochun Li, PhD, Co-Chair

Lingling Li, PhD

Doctoral Committee

Huiping Xu, PhD

December 5, 2016

Jennifer Wessel, PhD

DEDICATION

To My Family

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my advisor and mentor, Dr. Changyu Shen, for his guidance and advice throughout my Ph.D. program. I thank Dr. Xiaochun Li, Dr. Lingling Li, Dr. Huiping Xu, and Dr. Jennifer Wessel for serving in my research committee.

I am very grateful to my parents for their love, support and inspiration. They taught me to always be positive and optimistic.

I thank the entire Department of Biostatistics for creating a welcome, collaborative, but challenging atmosphere, which I truly enjoyed for the past five years. My gratitude goes out to all my friends, for their support, encouragement, and all the laughter they brought me especially during the occasional hardship throughout the Ph.D. program.

Jia Zhan

NEW STATISTICAL METHODS FOR THE EVALUATION OF EFFECTIVENESS AND SAFETY OF A MEDICAL INTERVENTION IN USING OBSERVATIONAL DATA

Observational studies offer unique advantages over randomized clinical trials (RCTs) in many situations where RCTs are not feasible or suffer from major limitations such as insufficient sample sizes and narrowly focused populations. Because observational data are relatively easy and inexpensive to access, and contain rich and comprehensive demographic and medical information on large and representative populations, they have played a major role in the assessment of the effectiveness and safety of medical interventions. However, observational data also have the challenges of higher rates of missing data and the confounding effect.

My proposal is on the development of three statistical methods to address these challenges. The first method is on the refinement and extension of a multiply-robust (MR) estimation procedure that simultaneously accounts for the confounding effect and missing covariate process, where we derived the asymptotic variance estimator and extended the method to the scenario where the missing covariate is continuous. The second method focuses on the improvement of estimation precision in an RCT by a historical control cohort. This was achieved through augmenting the conventional effect estimator with an extra mean zero (approximately) term correlated with the conventional effect estimator. In the third method, we calibrated the hidden database bias of an electronic medical records database and utilized an empirical Bayes method to improve the accuracy of the estimation of the risk of acute

myocardial infarction associated with a drug by borrowing information from other drugs.

Changyu Shen, PhD, Co-Chair

Xiaochun Li, PhD, Co-Chair

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

Interest in the use of electronic healthcare databases for comparative effectiveness and safety research is increasing [1-4]; these data are readily available, relatively inexpensive to access, and contain rich and comprehensive demographic and medical information on large, representative populations. In many scenarios in which clinical trials are not feasible due to either ethical or practical reasons, these observational databases provide valuable opportunities for the effectiveness and safety research. My dissertation contains three related topics of new statistical methods in using observational data.

In Chapter 2, I focus on a multiple robust (MR) method for missing data in observational study. In Chapter 3, I plan to use the historical data to improve the precision of estimation of the treatment effect from a small randomized control trial (RCT). In Chapter 4, by incorporating the existing knowledge, I utilize an empirical Bayes method for the evaluation of the drug safety from the observational data (electronic medical records (EMR) database).

1.1 A Multiply-robust Approach for Missing Data in Observational Study

Confounding bias and missing data are two major barriers to valid comparative effectiveness studies using observational electronic healthcare data. Each respective problem has been extensively studied. However, a unified approach to handle both issues simultaneously would be very helpful for observational studies with missing data problem. Multiply-robust (MR) methodology, which builds upon the well-established doubly-robust theory (proposed by Lingling Li, Williamson[5] independently), is such a remedy. The MR method is 4-fold robust in that it is consistent and asymptotic normal if one of four sets of modeling assumptions holds. However, the asymptotic variance of MR estimator hasn't been explored. In this chapter, we develop the asymptotic variance of MR estimator and extend the method from a missing binary covariate scenario to a missing continuous covariate scenario. We evaluate the performance of the MR estimate with a binary or continuous missing covariate and its asymptotic variance via a simulation study. The simulation study uses synthetic datasets to evaluate the performance of the MR method under a variety of settings, which cover most of the scenarios in the practical problems.

1.2 An Inverse Probability Weighting Approach for Historical Control Augmented Randomized Clinical Trials

The randomized control trial (RCT) is often considered as the gold standard for evaluating the effectiveness or safety of various types of medical interventions. While in the practical world, there might be difficulty in getting enough patients in to RCT especially for rare diseases. The remedy for this situation is to borrow information from

similar historical studies. To combine the historical data in to RCT, people now are mainly using Bayesian approach such as power prior and Bayesian hierarchical model. In this dissertation, we explored a frequentist approaching utilizing inverse probability weighting method to adjust for the difference in covariates between RCT and historical data. The IPW estimate could bring us more efficiency while still keep the unbiased property. We carried out a simulation to evaluate its performance compared with the estimate from RCT only.

1.3 The Associations of Drugs with Acute Myocardial Infarction: Bias Correction, Global Profiling and Inference on Individual Drug

We address two issues in drug-outcome association studies. First, it has been recognized that electronic health records databases may have database biases, for example, failure or incomplete capture of exposure and covariates, such that confounding cannot be fully controlled. In addition, there may be biases due to model misspecification. Consequently, risk estimates may be biased, resulting in the misguided assessment of the strength and direction of drug-outcome associations. Second, the distribution of the risk measures of relevant drugs on market for a given outcome not only characterizes the global profile of drug-outcome associations, but also can be used to improve the accuracy of the association estimate for future drugs through Bayes inference. Using acute myocardial infarction (AMI) as an example, we illustrate how the first issue can be addressed by calibrating the risk measures through drugs known to have no association with AMI in a population-level electronic medical records database, the

Indiana Network for Patient Care (INPC). We then employ an empirical Bayes approach to address the second issue.

The study shows that without the bias correction, 68.1%, 11.9% and 2.3% of the drugs included have a risk ratio for AMI greater than 1, 1.5 and 2, respectively. After the bias correction, the proportions become 51.5%, 7.0% and 1.4%, respectively. Using the empirical Bayes method, we gain 45% (without bias correction) and 52% (with bias correction) precision for the estimation of the risk ratio of a hypothetical new drug. Our approach serves as a general strategy for pharmaco-epidemiology studies for either an individual drug-outcome pair or multiple drug-outcome pairs.

CHAPTER 2. A MULTIPLY-ROBUST APPROACH FOR MISSING DATA IN OBSERVATIONAL STUDY

2.1 Background

Since observational databases are collected primarily for non-research purposes, they might have several limitations for evaluation of effectiveness and precision. First of all, confounding bias [6] exists and needs to be adjusted for when they are used to derive causal inference on treatments or interventions. Moreover, missing confounder and outcome data commonly occur due to various reasons [7]. Each respective problem has been extensively studied. Recently, there exists the multiply-robust (MR) approach [5], for missing outcome and confounder data in comparative effectiveness studies using observational databases. The method applies the double-robust (DR) method in a nested fashion to handle confounding bias and missing confounder data (either binary or continuous) in a unified manner. It is 4-fold robust [8] that it is consistent if one of four sets of modeling assumptions holds. [9-14]

In this chapter, we introduce the MR method and focus on the asymptotic characteristics of the MR estimates. In the Methods section, we introduce our notation, review the existing DR methods for causal inference and missing data models [12, 13], and present the MR method in dealing with either a binary or a continuous missing variables. For the purpose of comparison, we also introduce two non-MR competitors. Then we induce the asymptotic variance estimate of MR estimate using Taylor expansion. Next in the Simulation section, we describe the simulation studies to evaluate and compare the performance of the MR method with the two alternatives. We end with Discussion.

2.2 Methods

2.2.1 Notation

Let $\overline{\mathbf{O}} = \{\mathbf{O}_i = (Y_i, A_i, \mathbf{X}_i, R_i, R_i Z_i), i = 1, \dots, n\}$ denote n copies of independently and identically distributed data, where Y_i denotes subject i's outcome (continuous or binary), A_i denotes the binary treatment indicator (1 for intervention arm and 0 for control arm), X_i denotes a p-dimensional vector of potential confounders that are always observed, Z_i denotes the confounder (continuous or binary) that is subject to missingness, and R_i denotes the dichotomous missing indicator for Z_i, i.e., Z_i is observed if and only if $R_i = 1$. Further, let $\left(Y_{i,1}, Y_{i,0}\right)$ denote subject i's two potential outcomes for treatment and no treatment respectively. The parameter of interest is the average treatment effect, i.e., $\psi \equiv E\left[Y_{i,1}\right] - E[Y_{i,0}]$. Let ψ_0 denote the true value for ψ. The results below can be easily extended to other effect measures such as the causal relative risk $E\left[Y_{i,1}\right]/E\left[Y_{i,0}\right]$ for binary outcomes.

Throughout the paper, we impose the following assumptions.

Assumption 1. Consistency: $Y = Y_A$.

Assumption 2. Positivity: $\Pr(A = 1|X, Z) > 0$ with probability 1.

Assumption 3. No unmeasured confounders (NUC): (Y_1, Y_0) are conditionally independent of A given (X, Z).

Assumption 4. Missing at random (MAR): $\Pr(R = 1|Y, A, X, Z) = \Pr(R = 1|Y, A, X)$ and $\Pr(R = 1|Y, A, X) > 0$ with probability 1.

Let $b(A, X, Z)$ denote $E[Y|A, X, Z]$, the outcome regression (OR) function; and let $e(X, Z)$ denote $\Pr(A = 1|X, Z)$, the propensity score (PS) function which is defined as

the conditional probability of receiving treatment given the confounders. Under NUC, $(Y_1, Y_0) \perp A|e(X,Z)$ [11]. Let $\pi(Y,A,X)$ denote $\Pr(R=1|Y,A,X,Z) = \Pr(R=1|Y,A,X)$ under MAR, the conditional probability of observing complete data. When Z is binary, let $\varpi(Y,A,X)$ denote $\Pr(Z=1|Y,A,X) = \Pr(Z=1|R=1,Y,A,X)$ under MAR, the conditional probability of having $Z=1$ given (Y,A,X) among subjects with complete data. When Z is continuous, let $\varpi(Y,A,X)$ denote $\mathrm{E}(Z|Y,A,X) = \mathrm{E}(Z|R=1,Y,A,X)$ under MAR, the conditional expectation of Z given (Y,A,X) among subjects with complete data. Let $P_n\{h(O)\} = \frac{1}{n}\sum_{i=1}^{n} h(O)$ denote the sample mean of any given function $h(\cdot)$.

2.2.2 DR estimation in causal inference models

We first consider a hypothetical causal inference model in which complete confounder data is observed for all study subjects, i.e., data is denoted by $\overline{\mathbf{O}}^F = \{\mathbf{O}_i^F = (Y_i, A_i, \mathbf{X}_i, Z_i),\ i = 1, \dots, n\}$. Under the NUC assumption, there are in general two classes of methods that are commonly used to adjust for confounding, one based on the OR model, the other on the PS model. The OR-based analysis depends on the OR function $b(A, \mathbf{X}, Z)$. A parametric working model, $b(A, \mathbf{X}, Z; \eta)$, is imposed for $b(A, \mathbf{X}, Z)$ and fitted to data $\{(Y_i, A_i, \mathbf{X}_i, Z_i), i = 1, \dots, n\}$ to obtain the maximum likelihood estimator of η, $\tilde{\eta}$. Then the OR-based estimator of ψ, $\tilde{\psi}_{OR}$, can be calculated as $\mathrm{P}_n\{b(A=1, \mathbf{X}, Z; \tilde{\eta})\} - \mathrm{P}_n\{b(A=0, \mathbf{X}, Z; \tilde{\eta})\}$. The OR-based estimator $\tilde{\psi}_{OR}$ is a consistent estimator of ψ if the imposed parametric working model $b(A, \mathbf{X}, Z; \eta)$ is correct. Standard regression methods such as linear and logistic regressions belong to this class. The PS-based analysis depends on the PS function $e(\mathbf{X}, Z)$. A parametric working model $e(\mathbf{X}, Z; \alpha)$ (e.g., a logistic regression model) is imposed for $e(\mathbf{X}, Z)$ and fitted to data

$\{(A_i, \mathbf{X}_i, Z_i), i = 1, \dots, n\}$ to obtain the maximum likelihood estimator of α, $\tilde{\alpha}$, and the estimated PSs $\{\tilde{e}_i \equiv e(\mathbf{X}_i, Z_i; \tilde{\alpha}), i = 1, \dots, n\}$. PSs can be used in different ways: for matching [11], stratification [10], regression adjustment [9], and weighting [9, 12, 13]. For instance, a PS weighting estimator

$\tilde{\psi}_{IPW} \equiv \left(P_n\left\{\frac{I(A=1)}{e(\mathbf{X},Z;\tilde{\alpha})}\right\}\right)^{-1} \mathrm{P}_n\left\{\frac{I(A=1)}{e(\mathbf{X},Z;\tilde{\alpha})}Y\right\} - \left(P_n\left\{\frac{I(A=0)}{1-e(\mathbf{X},Z;\tilde{\alpha})}\right\}\right)^{-1} \mathrm{P}_n\left\{\frac{I(A=0)}{1-e(\mathbf{X},Z;\tilde{\alpha})}Y\right\}$. The validity of the PS-based methods requires the PS model is correctly specified.

The two classes of approaches have their own respective merits. Nevertheless, the OR-based methods require the OR function $b(A, \mathbf{X}, Z)$ to be correctly specified, and the PS-based methods require the PS function $e(\mathbf{X}, Z)$ to be correctly specified. Model misspecification remains a threat to the validity of analyses results no matter which method is chosen. The DR method combines both OR and PS models in a fashion that gives valid inference if either model is correct, but not necessarily both [14]. Thus, it offers dual protection against model misspecifications. Furthermore, even when both the OR and PS models are misspecified (which is not uncommon in practice with many covariates and complex function forms), if either model is nearly correct, the DR estimator, according to theory and simulation results [14, 15], substantially reduces the bias. Thus, this method offers the analyst two chances to make nearly correct inference on the parameter of interest. Specifically, the DR estimator of ψ, $\psi_{DR}(\overline{\mathbf{O}}^F; \alpha, \eta)$, is obtained by solving the estimating equation below

$$\mathrm{P}_n\{\mu_1^F(\mathbf{O}^F; \alpha, \eta) - \mu_0^F(\mathbf{O}^F; \alpha, \eta)\} - \psi = 0, \tag{1}$$

where

$$\mu_1^F(\mathbf{O}^F; \alpha, \eta) = \frac{I(A=1)}{e(\mathbf{X}, Z; \alpha)}Y - \left(\frac{I(A=1)}{e(\mathbf{X}, Z; \alpha)} - 1\right) b(A = 1, \mathbf{X}, Z; \eta), \tag{2}$$

$$\mu_0^F(\mathbf{O}^F;\alpha,\eta) = \frac{I(A=0)}{1-e(\mathbf{X},Z;\alpha)}Y - \left(\frac{I(A=0)}{1-e(\mathbf{X},Z;\alpha)} - 1\right) b(A=0,\mathbf{X},Z;\eta). \quad (3)$$

Let $\tilde{\psi}_{DR}$ denote $\psi_{DR}(\overline{\mathbf{O}}^F;\tilde{\alpha},\tilde{\eta})$. It is known that $\tilde{\psi}_{DR}$ is a consistent and asymptotically normal (CAN) estimator of ψ if either of the OR and PS working models is correct, but not necessarily both [14].

2.2.3 DR estimation in missing data

The DR method in missing data models shares the same heuristic ideas as that in causal inference models. In fact, the causal inference models can be conceptually viewed as missing data models. The average treatment effect ψ is the difference between the marginal means of the two potential outcomes. For the marginal mean $E[Y_a], a \in \{0,1\}$, the potential outcome Y_a is observed only for a subset of subjects with $A_i = a$. Thus, in this conceptual "missing data" model, the potential outcome Y_a is the component that is subject to missingness and the dummy variable $I(A_i = a)$ is the "missing-data indicator".

Consider a general missing data model in which R_i denotes subject i's missing indicator, $\mathbf{O}_i^F$ denotes the full data and $\mathbf{O}_i$ denotes the observed data, and $\theta = E[h(\mathbf{O}_i^F)]$ denotes the parameter of interest, the marginal mean of a known function $h(\cdot)$. We assume the MAR assumption holds, i.e., $\Pr(R=1|\mathbf{O}_i^F) = \Pr(R=1|\mathbf{O}_i)$, which is denoted as $\pi(\mathbf{O}_i)$.

Based on the DR theory for missing data, it can be shown that the DR estimator of θ, $\theta_{DR}\left(\overline{\mathbf{O}} = (\mathbf{O}_1,\dots,\mathbf{O}_n)\right)$, is obtained by solving the estimating equation below.

$$\mathrm{P}_n\left\{\frac{R}{\pi(\mathbf{O}_i)}h(\mathbf{O}_i^F) - \left(\frac{R}{\pi(\mathbf{O}_i)} - 1\right)E[h(\mathbf{O}_i^F)|R=1,\mathbf{O}_i]\right\} - \theta = 0. \quad (4)$$

Heuristically, this approach weights subjects who have complete data by $\pi^{-1}(\mathbf{O}_i)$ to remove the selection bias due to missing data, and adds an augmentation term $\left(\frac{R}{\pi(\mathbf{O}_i)} - 1\right) E[h(\mathbf{O}_i^F)|R = 1, \mathbf{O}_i]$ to increase efficiency and provide the double-robustness property.

2.2.4 MR estimation to handle missing confounder in causal inference models

Next, we introduce the MR estimator of ψ, $\hat{\psi}_{MR}$, that depends on the observed data $\overline{\mathbf{O}}$ only and has a multiple-robustness bias property that will be discussed later this section.

<u>2.2.4.1 The MR estimator</u>

The MR method uses DR estimation in causal inference and missing data models in a nested fashion, and constructs the estimating function for ψ_{MR} in two stages. The first stage is DR estimation in a hypothetical causal inference model with complete confounder data $\overline{\mathbf{O}}^F = \{\mathbf{O}_i^F = (Y_i, A_i, \mathbf{X}_i, Z_i), i = 1, \dots, n\}$. Note that the estimating function for $\psi_{DR}(\overline{\mathbf{O}}^F; \alpha, \eta)$ is $\{\mu_1^F(\mathbf{O}^F; \alpha, \eta) - \mu_0^F(\mathbf{O}^F; \alpha, \eta) - \psi\}$ with $\mu_1^F(\mathbf{O}^F; \alpha, \eta)$ and $\mu_0^F(\mathbf{O}^F; \alpha, \eta)$ defined in Eqs. (2) and (3) respectively. This estimating function depends on the confounder Z, which, in the observed data $\overline{\mathbf{O}} = \{\mathbf{O}_i = (Y_i, A_i, \mathbf{X}_i, R_i, R_i Z_i), i = 1, ., n\}$, is available only for the subset of subjects with $R_i = 1$. To address this issue, in the second stage, we apply DR estimation to estimate $\theta = E[h^*(\mathbf{O}_i^F; \alpha, \eta)]$ where $h^*(\mathbf{O}_i^F; \alpha, \eta) \equiv \mu_1^F(\mathbf{O}^F; \alpha, \eta) - \mu_0^F(\mathbf{O}^F; \alpha, \eta)$. Let $\pi(Y, A, \mathbf{X}; \gamma)$ denote the imposed parametric working model for $\pi(Y, A, \mathbf{X}) \equiv \Pr(R = 1|Y, A, \mathbf{X}, Z) = \Pr(R = 1|Y, A, \mathbf{X})$ under MAR, and $\varpi(Y, A, \mathbf{X}; \omega)$ denote the imposed parametric working model for $\varpi(Y, A, \mathbf{X}) \equiv \Pr(Z = 1|Y, A, \mathbf{X}) = \Pr(Z = 1|R = 1, Y, A, \mathbf{X})$ under MAR. Then by Eq.

(4), the estimating function for $\theta_{DR}(\overline{\mathbf{O}}, \alpha, \eta, \gamma, \omega) = \frac{R}{\pi(Y,A,\mathbf{X};\gamma)} h^*(\mathbf{O}^F; \alpha, \eta) - \left(\frac{R}{\pi(Y,A,\mathbf{X};\gamma)} - 1\right) E[h^*(\mathbf{O}^F; \alpha, \eta) | R = 1, Y, A, \mathbf{X}; \omega)] - \theta$. After plugging in the expressions of $\mu_1^F(\mathbf{O}^F; \alpha, \eta)$ and $\mu_0^F(\mathbf{O}^F; \alpha, \eta)$, we obtain the estimating function for $\psi_{MR}(\overline{\mathbf{O}}; \alpha, \eta, \gamma, \omega)$ as

$$IF_{MR}(\overline{\mathbf{O}}; \alpha, \eta, \gamma, \omega, \psi)$$

$$\equiv \frac{R}{\pi(Y, A, \mathbf{X}; \gamma)} \times \left\{ \begin{array}{c} \left(\frac{I(A=1)}{e(\mathbf{X}, Z; \alpha)} Y - \left(\frac{I(A=1)}{e(\mathbf{X}, Z; \alpha)} - 1 \right) b(A = 1, \mathbf{X}, Z; \eta) \right) - \\ \left(\frac{I(A=0)}{1 - e(\mathbf{X}, Z; \alpha)} Y - \left(\frac{I(A=0)}{1 - e(\mathbf{X}, Z; \alpha)} - 1 \right) b(A = 0, \mathbf{X}, Z; \eta) \right) \end{array} \right\}$$

$$- \left(\frac{R}{\pi(Y, A, \mathbf{X}; \gamma)} - 1 \right)$$

$$\times E \left[\left\{ \begin{array}{c} \left(\frac{I(A=1)}{e(\mathbf{X}, Z; \alpha)} Y - \left(\frac{I(A=1)}{e(\mathbf{X}, Z; \alpha)} - 1 \right) b(A = 1, \mathbf{X}, Z; \eta) \right) - \\ \left(\frac{I(A=0)}{1 - e(\mathbf{X}, Z; \alpha)} Y - \left(\frac{I(A=0)}{1 - e(\mathbf{X}, Z; \alpha)} - 1 \right) b(A = 0, \mathbf{X}, Z; \eta) \right) \end{array} \right\} \middle| R = 1, Y, A, \mathbf{X}; \omega \right]$$

$$-\psi.$$

It is obvious that $IF_{MR}(\overline{\mathbf{O}}; \alpha, \eta, \gamma, \omega, \psi)$ depends on observed data $\overline{\mathbf{O}}$, as well as 4 nuisance parameters $(\alpha, \eta, \gamma, \omega)$. The MLE of γ, $\hat{\gamma}$, can be easily obtained by fitting the imposed parametric model $\pi(Y, A, \mathbf{X}; \gamma)$ for $\Pr(R_i = 1 | Y_i, A_i, \mathbf{X}_i)$ on the observed data $\{(R_i, Y_i, A_i, \mathbf{X}_i), i = 1, \ldots n\}$. The most commonly used working models for $\pi(Y, A, \mathbf{X})$ are logistic regression models.

- When Z is binary, for any given function $h(Y, A, \mathbf{X}, Z)$,

$$E[h(Y, A, \mathbf{X}, Z) | R = 1, Y, A, \mathbf{X}; \omega]$$

$$= h(Y, A, \mathbf{X}, Z = 1) \varpi(Y, A, \mathbf{X}; \omega) + h(Y, A, \mathbf{X}, Z = 0) \big(1 - \varpi(Y, A, \mathbf{X}; \omega)\big).$$

The MLE of ω, $\widehat{\omega}$, can be obtained by fitting the imposed parametric model $\varpi(Y, A, \mathbf{X}; \omega)$ for $Pr(Z_i = 1|R_i = 1, Y_i, A_i, \mathbf{X}_i)$ on the observed data $\{(Z_i, Y_i, A_i, \mathbf{X}_i)|i = 1, \ldots n \; and \; R_i = 1\}$ among the subset of complete cases.

- When Z is continuous,

$$E[h(Y, A, \mathbf{X}, Z)|R = 1, Y, A, \mathbf{X}; \omega] \cong h(Y, A, \mathbf{X}, \widehat{\omega}).$$

Since Z is continuous, the expectation is obtained through integration while could be computationally demanding. Here we use $\widehat{\omega}$, the MLE of ω, obtained by fitting the imposed parametric model $\varpi(Y, A, \mathbf{X}; \omega)$ for $E(Z_i|R_i = 1, Y_i, A_i, \mathbf{X}_i)$ on the observed data $\{(Z_i, Y_i, A_i, \mathbf{X}_i)|i = 1, \ldots n \; and \; R_i = 1\}$ among the subset of complete cases, to approximate the expectation of h.

The estimation of α and η is less straightforward since their MLEs depend on $\{Z_i, i = 1, \ldots, n\}$, which is only partially observed in $\overline{\mathbf{O}}$. For instance, $\hat{\alpha}$ is typically obtained by solving $\mathrm{P}_n\{S_\alpha(A, \mathbf{X}, Z; \alpha)\} = 0$ where $S_\alpha(A, \mathbf{X}, Z; \alpha)$ is the score function for α and equals $\big(A - e(\mathbf{X}, Z; \alpha)\big)(1, X^T, Z)^T$ with a logistic working model $logit(e(\mathbf{X}, Z; \alpha)) = (1, \mathbf{X}^T, Z)\alpha$. But in $\overline{\mathbf{O}}$, Z_i is observed if and only if $R_i = 1$. To address this issue, we apply DR estimation to the estimating function of $\hat{\alpha}$, and obtain a DR estimator of α, $\hat{\alpha}_D(\hat{\gamma}, \widehat{\omega})$, by solving the estimating equation below

$$\mathrm{P}_n \left\{ \begin{array}{c} \dfrac{R}{\pi(Y, A, \mathbf{X}; \hat{\gamma})} S_\alpha(A, \mathbf{X}, Z; \alpha) - \left(\dfrac{R}{\pi(Y, A, \mathbf{X}; \hat{\gamma})} - 1 \right) \times \\ E[S_\alpha(A, \mathbf{X}, Z; \alpha)|R = 1, Y, A, \mathbf{X}; \widehat{\omega}] \end{array} \right\} = 0 \qquad (5)$$

Suppose α_0 denotes the solution to $E[S_\alpha(A, \mathbf{X}, Z; \alpha)] = 0$. The estimator $\hat{\alpha}_D(\hat{\gamma}, \widehat{\omega})$ is DR in the sense that it is a consistent estimator of α_0 if either the working model $\pi(Y, A, \mathbf{X}; \gamma)$ or the working model $\varpi(Y, A, \mathbf{X}; \omega)$ is correctly specified, but not necessarily both. Similarly, we construct the DR estimator of η, $\hat{\eta}_D(\hat{\gamma}, \widehat{\omega})$, by applying DR estimation to

the estimating function of $\hat{\eta}$. Specifically, $\hat{\eta}_D(\hat{\gamma}, \hat{\omega})$ is obtained by solving the estimating equation below

$$\mathrm{P}_n \left\{ \begin{array}{c} \frac{R}{\pi(Y, A, \mathbf{X}; \hat{\gamma})} S_\eta(Y, A, \mathbf{X}, Z; \eta) - \left(\frac{R}{\pi(Y, A, \mathbf{X}; \hat{\gamma})} - 1 \right) \\ \times E\left[S_\eta(Y, A, \mathbf{X}, Z; \eta) | R = 1, Y, A, \mathbf{X}; \hat{\omega} \right] \end{array} \right\} = 0, \tag{6}$$

where $S_\eta(Y, A, \mathbf{X}, Z; \eta)$ is the score function for η.

Finally, we have the MR estimator

$$\hat{\psi}_{MR} \equiv \psi_{MR}(\bar{\mathbf{O}}; \hat{\alpha}_D(\hat{\gamma}, \hat{\omega}), \hat{\eta}_D(\hat{\gamma}, \hat{\omega}), \hat{\gamma}, \hat{\omega}).$$

<u>2.2.4.2 Asymptotic variance of $\hat{\psi}_{MR}$</u>

The MR estimator $\hat{\psi}_{MR}$ depends on the following four working models:

- Model (i): $b(A, \mathbf{X}, Z; \eta)$ for the OR function $E[Y|A, \mathbf{X}, Z]$
- Model (ii): $e(\mathbf{X}, Z; \alpha)$ for the PS function $\Pr(A = 1|\mathbf{X}, Z)$
- Model (iii): $\pi(Y, A, \mathbf{X}; \gamma)$ for the missing indicator function $\Pr(R = 1|Y, A, \mathbf{X})$
- Model (iv): $\varpi(Y, A, \mathbf{X}; \omega)$ for the conditional mean function $\Pr(Z = 1|R = 1, Y, A, \mathbf{X})$ (binary Z) or $E(Z|R = 1, Y, A, \mathbf{X})$ (continuous Z)

Under mild regularity conditions [16] and Assumptions 1-4, $\hat{\psi}_{MR}$ is a CAN estimator of ψ_0 if at least one of the following four conditions holds:

- Condition 1: working model (i) for $E[Y|A, \mathbf{X}, Z]$ and model (iii) for $\Pr(R = 1|Y, A, \mathbf{X})$ are correct
- Condition 2: working model (ii) for $\Pr(A = 1|\mathbf{X}, Z)$ and model (iii) for $\Pr(R = 1|Y, A, \mathbf{X})$ are correct
- Condition 3: working model (i) for $E[Y|A, \mathbf{X}, Z]$ and model (iv) for $\Pr(Z = 1|R = 1, Y, A, \mathbf{X})$ (binary Z) or $E(Z|R = 1, Y, A, \mathbf{X})$ (continuous Z) are correct

- Condition 4: working model (ii) for $\Pr(A = 1|\mathbf{X}, Z)$ and model (iv) for $\Pr(Z = 1|R = 1, Y, A, \boldsymbol{X})$ (binary Z) or $E(Z|R = 1, Y, A, \mathbf{X})$ (continuous Z) are correct.

We provide the proof of asymptotic normality and consistency, the induction of the asymptotic variance $Var(\hat{\psi}_{MR})$ in Appendix A. More details about the variance formula for binary/continuous missing covariate Z are elaborated in Appendix B An empirical VARIANCE ESTIMATe of $\hat{\varphi}_a$ and Appendix C An empirical VARIANCE ESTIMATe of $\hat{\varphi}_a$.

2.2.5 Two alternative approaches

For the purpose of comparison, we introduce two alternative approaches.

<u>2.2.5.1 The complete-case approach</u>

The complete-case approach is the simplest and possibly the most commonly used approach to deal with missing data. It conducts standard analyses among complete cases, i.e., the subset of subjects with complete data on relevant variables. In our setting, the complete-case approach fits the imposed working model $b(A, \mathbf{X}, Z; \eta)$ for $E[Y|A, \mathbf{X}, Z]$ among the complete cases $(R = 1)$ to obtains an MLE of η, $\check{\eta}$; uses the fitted model to predict outcome means $b(A = 1, \mathbf{X}, Z; \check{\eta})$ and $b(A = 0, \mathbf{X}, Z; \check{\eta})$ for each complete case; and finally estimates ψ by taking the sample average of the mean difference among complete cases. Specifically,

$$\hat{\psi}_{CC} = (\mathrm{P}_n\{I(R = 1)\})^{-1}\mathrm{P}_n\{I(R = 1)[b(A = 1, \mathbf{X}, Z; \check{\eta}) - b(A = 0, \mathbf{X}, Z; \check{\eta})]\}.$$

Even when 1) the working model $b(A, \mathbf{X}, Z; \eta)$ is correct in that $E[Y|A, \mathbf{X}, Z] = b(A, \mathbf{X}, Z; \eta_0)$, and 2) the MAR assumption holds, the complete-case approach may still be biased as $E[Y|A, \mathbf{X}, Z]$ may differ from $E[Y|R = 1, A, \mathbf{X}, Z]$ and thus $\check{\eta}$ is not a

consistent estimator for η_0. If the missing data probability $\Pr(R = 1|Y, A, \mathbf{X}, Z)$ depends on the observed data through $(A, \mathbf{X})$ only, then it can be easily shown that the complete-case approach is valid when models (i) is correct and the MAR assumption holds.

2.2.5.2 The full data approach

Full data is the ideal case without missing data (all Z observed) $\hat{\psi}_{FD}$. Full likelihood approach should give out the best estimation results compared with MR and complete cases.

There exist other alternative estimators that are consistent for ψ under other conditions. For instance, we may use regression to handle confounding bias and then weighting to handle missing data. Or we may use DR to handle confounding and then weighting to handle missing data. They are not discussed further in this chapter for space constraints.

2.3 Simulation

We conducted a simulation study to evaluate the performance of the MR estimator $\hat{\psi}_{MR}$ in the presence of varying degree of selection and confounding bias, and compared with the alternative estimators $\hat{\psi}_{CC}$ and $\hat{\psi}_{FD}$ in bias, standard error, and coverage probability of the 95% CI.

We considered the data generating mechanisms as following:

- We first generated $\mathbf{X} = (X_1, \cdots, X_5)$
  - X_1 follows a mixed uniform distribution $0.35\mathrm{U}[40,50] + 0.32\mathrm{U}[50,60] + 0.2\mathrm{U}[60,70] + 0.07\mathrm{U}[70,80] + 0.06\mathrm{U}[80,90]$, and then is rounded to integer for the practical meaning of age.
  - X_2 follows Bernoulli distribution with a probability 0.5.

- X_3 follows the multinomial distribution depending on X_1. (See Table 1)

- X_4 follows the standard normal distribution.

Table 1 Marginal distribution of X_3

X_1	$Pr(X_3 = 0)$	$Pr(X_3 = 1)$	$Pr(X_3 = 2)$
≤ 44	0.390	0.297	0.313
$[45, 64]$	0.266	0.350	0.384
$[65, 74]$	0.261	0.391	0.348
≥ 75	0.362	0.397	0.241

Let $\beta_X = (\log 1.5, \log 1.5, \log 1.5, \log 1.5)^T$ be the coefficient of $\mathbf{X}$.

- We then generated the binary confounder Z based on a conditional mean function $\Pr(Z|\mathbf{X}) = \frac{1}{1+\exp(-\beta_{0,z}-\beta_x\mathbf{X})}$, or the continuous confounder Z based on $E(Z|\mathbf{X}) = \frac{1}{1+\exp(-\beta_{0,z}-\beta_x\mathbf{X})}$.

- We then generated the binary exposure A based on the PS function $(\mathbf{X}, Z) = \frac{1}{1+\exp(-\beta_{0,a}-\log 1.5\times Z-\beta_x\mathbf{X})}$.

- We then generated the binary outcome Y based on a conditional mean function $b(A, \mathbf{X}, Z) = \frac{1}{1+\exp(-\beta_{0,y}-\log 1.5\times Z-\beta_x\mathbf{X}-\log(2)\times A)}$

- Finally we generated the missing indicator R based on a conditional mean function $\pi(Y, A, \mathbf{X}) = \frac{1}{1+\exp(-\beta_{0,r}-\beta_x\mathbf{X}-\log(1.5)A-\log(1.8)Y)}$ and set the value of Z to missing if $R = 0$.

The coefficients $(\beta_{0,z}, \beta_{0,a}, \beta_{0,y}, \beta_{0,r})$ were determined such that the marginal means of Y, A, R, Z were fixed around 0.1, 0.5, 0.6, and 0.4 (baseline) respectively. Table 2 presents the marginal mean of other settings.

Table 2 Settings for marginal means of $Z; A; Y; R$

Setting	Z	A	Y	R
1(Baseline)	0.4	0.5	0.1	0.6
2,3	0.2/0.8	0.5	0.1	0.6
4,5	0.4	0.2/0.8	0.1	0.6
6,7	0.4	0.5	0.05/0.1	0.6
8	0.4	0.5	0.1	0.3

In each setting, we conducted 1000 Monte Carlo (MC) trials. We obtained both the estimated asymptotic variance and the MC variance of $\hat{\psi}_{MR}$. We construct the 95% Wald CI using the estimated asymptotic variance. To calculate bias we needed to know the true value ψ_0 of ψ, we obtained ψ_0 (with negligible error) by taking the empirical average of $[b(A = 1, \mathbf{X}, Z) - b(A = 0, \mathbf{X}, Z)]$ from a million observations available in the 1000 MC trials. In fitting the data, when we wish to construct a mis-specified working model for a particular one of the above 4 functions we simply (incorrectly) set β_x to $(0, \log 1.5\,, \log 1.5\,, \log 1.5)^T$ in the working model.

We present the results for the four considered estimators, $\hat{\psi}_{CC}$, $\hat{\psi}_{REG}$ and $\hat{\psi}_{MR}$ across all simulation settings and within each setting, under the following 6 different scenarios:

Scenario 1: working models (i)-(iii) are correct, working model (iv) is "correct" (in the sense explained below)

Scenario 2: working models (i)-(iv) are incorrect

Scenario 3: working models (i) and (iii) are correct, working models (ii) and (iv) are incorrect

Scenario 4: working models (ii) and (iii) are correct, working models (i) and (iv) are incorrect

Scenario 5: working model (i) is correct, working model (iv) is correct, working models (ii) and (iii) are incorrect

Scenario 6: working model (ii) is correct, working model (iv) is correct, working models (i) and (iii) are incorrect

The complete-case estimator $\hat{\psi}_{CC}$ varies only with the working model (i) for $E[Y|A, \mathbf{X}, Z]$. It is not guaranteed to be consistent for ψ_0 even when model (i) is correct because $E[Y|A, \mathbf{X}, Z]$ differs from $E[Y|R = 1, A, \mathbf{X}, Z]$ in the considered settings. The MR estimator $\hat{\psi}_{MR}$ varies across all 6 scenarios. It should be consistent in scenarios 1, 3, and 4. It might be slightly biased in scenarios 5 and 6 with only a correct model (iv).

From simulation, we get Table 3 and Table 4 for missing binary/continuous covariate Z under the baseline setting.

Table 3 Baseline results for missing binary covariate Z

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.810	0.564				
	FD	0.959	0.901				
	MR	0.946	0.888	0.942	0.947	0.944	0.946
Bias	CC	0.022	0.035				
	FD	-0.001	0.013				
	MR	-0.001	0.013	-0.002	-0.001	-0.002	-0.001
MC SE	CC	0.023	0.020				
	FD	0.021	0.020				
	MR	0.022	0.021	0.022	0.022	0.022	0.022
Average Analytic SE	CC	0.024	0.021				
	FD	0.021	0.021				
	MR	0.022	0.021	0.022	0.022	0.022	0.022

Table 4 Baseline results for missing continuous covariate Z

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.859	0.758				
	FD	0.937	0.875				
	MR	0.929	0.885	0.933	0.935	0.934	0.930
Bias	CC	0.017	0.030				
	FD	0.001	0.015				
	MR	0.001	0.015	0.001	0.001	0.002	0.001
MC SE	CC	0.028	0.027				
	FD	0.022	0.021				
	MR	0.023	0.022	0.022	0.023	0.022	0.023
Average Analytic SE	CC	0.027	0.026				
	FD	0.021	0.020				
	MR	0.022	0.021	0.022	0.022	0.022	0.022

For both baseline settings, we could find the MR estimate $\hat{\psi}_{MR}$ gives us acceptable bias and its coverage probabilities based on the empirical asymptotic variance are close to nominal level in most of the scenarios except for all models incorrect. The results also shows the MC SE is close to the average asymptotic variance estimate of $\hat{\psi}_{MR}$, which verifies the correctness of estimation of asymptotic variance of $\hat{\psi}_{MR}$. Compared with $\hat{\psi}_{CC}$ under scenario 1, the performance of $\hat{\psi}_{MR}$ is much better (higher coverage probability, smaller bias and smaller estimated asymptotic variance); while compared with $\hat{\psi}_{FD}$, $\hat{\psi}_{MR}$'s bias and coverage probability is very close to $\hat{\psi}_{FD}$'s under scenario 1 and 2.

For the setting 2~8, we summarize the results in the following tables, 5 through 18.

Table 5 Results for missing binary covariate Z in setting 2

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.868	0.680				
	FD	0.949	0.899				
	MR	0.939	0.891	0.942	0.943	0.940	0.940
Bias	CC	0.017	0.031				
	FD	-0.001	0.013				
	MR	0.000	0.013	-0.002	-0.001	-0.002	-0.000
MC SE	CC	0.025	0.022				
	FD	0.021	0.020				
	MR	0.022	0.021	0.022	0.023	0.022	0.023
Average Analytic SE	CC	0.024	0.022				
	FD	0.021	0.020				
	MR	0.021	0.021	0.021	0.022	0.021	0.022

Table 6 Results for missing binary covariate Z in setting 3

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.658	0.427				
	FD	0.968	0.894				
	MR	0.950	0.889	0.952	0.951	0.951	0.952
Bias	CC	0.027	0.039				
	FD	0.000	0.014				
	MR	0.000	0.014	0.000	0.000	0.000	0.000
MC SE	CC	0.025	0.020				
	FD	0.021	0.020				
	MR	0.021	0.021	0.021	0.021	0.021	0.021
Average Analytic SE	CC	0.024	0.019				
	FD	0.025	0.023				
	MR	0.021	0.020	0.021	0.021	0.021	0.021

Table 7 Results for missing binary covariate Z in setting 4

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.818	0.600				
	FD	0.961	0.937				
	MR	0.940	0.938	0.930	0.942	0.928	0.944
Bias	CC	0.018	0.029				
	FD	0.001	0.011				
	MR	-0.001	0.011	0.000	0.000	0.000	0.000
MC SE	CC	0.019	0.018				
	FD	0.018	0.019				
	MR	0.035	0.022	0.022	0.028	0.021	0.021
Average Analytic SE	CC	0.019	0.018				
	FD	0.019	0.020				
	MR	0.021	0.021	0.021	0.021	0.020	0.020

Table 8 Results for missing binary covariate Z in setting 5

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.825	0.678				
	FD	0.940	0.886				
	MR	0.920	0.879	0.919	0.921	0.917	0.922
Bias	CC	0.028	0.043				
	FD	0.002	0.019				
	MR	0.003	0.020	0.002	0.003	0.001	0.003
MC SE	CC	0.041	0.036				
	FD	0.032	0.030				
	MR	0.041	0.037	0.039	0.041	0.039	0.042
Average Analytic SE	CC	0.040	0.034				
	FD	0.033	0.031				
	MR	0.040	0.036	0.038	0.040	0.038	0.041

Table 9 Results for missing binary covariate Z in setting 6

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.820	0.650				
	FD	0.951	0.916				
	MR	0.936	0.905	0.943	0.931	0.940	0.933
Bias	CC	0.012	0.020				
	FD	-0.001	0.007				
	MR	-0.001	0.007	-0.001	-0.001	-0.001	-0.001
MC SE	CC	0.018	0.015				
	FD	0.016	0.016				
	MR	0.017	0.016	0.017	0.017	0.017	0.017
Average Analytic SE	CC	0.018	0.015				
	FD	0.017	0.016				
	MR	0.017	0.016	0.017	0.017	0.017	0.017

Table 10 Results for missing binary covariate Z in setting 7

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.813	0.542				
	FD	0.954	0.875				
	MR	0.946	0.880	0.943	0.931	0.940	0.933
Bias	CC	0.027	0.046				
	FD	0.000	0.018				
	MR	0.000	0.018	-0.001	0.000	-0.001	0.000
MC SE	CC	0.028	0.025				
	FD	0.024	0.024				
	MR	0.025	0.025	0.017	0.017	0.017	0.017
Average Analytic SE	CC	0.028	0.025				
	FD	0.024	0.024				
	MR	0.025	0.024	0.017	0.017	0.017	0.017

Table 11 Results for missing binary covariate Z in setting 8

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.789	0.667				
	FD	0.965	0.905				
	MR	0.961	0.900	0.946	0.958	0.943	0.957
Bias	CC	0.025	0.038				
	FD	0.000	0.014				
	MR	0.000	0.014	-0.001	0.000	-0.001	0.000
MC SE	CC	0.035	0.030				
	FD	0.021	0.020				
	MR	0.023	0.022	0.023	0.023	0.023	0.023
Average Analytic SE	CC	0.035	0.030				
	FD	0.023	0.022				
	MR	0.023	0.022	0.023	0.023	0.023	0.023

Table 12 Results for missing continuous covariate Z in setting 2

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.894	0.787				
	FD	0.944	0.901				
	MR	0.946	0.910	0.942	0.949	0.943	0.947
Bias	CC	0.014	0.027				
	FD	-0.001	0.013				
	MR	-0.001	0.013	-0.001	-0.001	-0.001	-0.001
MC SE	CC	0.028	0.026				
	FD	0.021	0.020				
	MR	0.022	0.021	0.022	0.022	0.022	0.022
Average Analytic SE	CC	0.027	0.026				
	FD	0.021	0.020				
	MR	0.022	0.021	0.021	0.022	0.021	0.022

Table 13 Results for missing continuous covariate Z in setting 3

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.893	0.788				
	FD	0.950	0.884				
	MR	0.946	0.876	0.943	0.949	0.942	0.948
Bias	CC	0.015	0.028				
	FD	0.000	0.014				
	MR	0.000	0.014	0.000	0.000	0.000	0.000
MC SE	CC	0.027	0.026				
	FD	0.021	0.021				
	MR	0.022	0.021	0.022	0.022	0.022	0.022
Average Analytic SE	CC	0.027	0.026				
	FD	0.021	0.020				
	MR	0.021	0.021	0.021	0.022	0.021	0.022

Table 14 Results for missing continuous covariate Z in setting 4

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.920	0.845				
	FD	0.944	0.928				
	MR	0.949	0.938	0.945	0.947	0.945	0.950
Bias	CC	0.012	0.023				
	FD	0.000	0.010				
	MR	0.000	0.010	0.000	0.000	0.000	0.000
MC SE	CC	0.023	0.023				
	FD	0.019	0.020				
	MR	0.021	0.022	0.021	0.021	0.021	0.021
Average Analytic SE	CC	0.022	0.023				
	FD	0.019	0.020				
	MR	0.020	0.021	0.020	0.020	0.020	0.020

Table 15 Results for missing continuous covariate Z in setting 5

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.866	0.802				
	FD	0.921	0.881				
	MR	0.925	0.898	0.930	0.922	0.930	0.924
Bias	CC	0.018	0.034				
	FD	0.000	0.017				
	MR	-0.001	0.017	-0.001	-0.001	0.000	0.000
MC SE	CC	0.048	0.044				
	FD	0.035	0.032				
	MR	0.041	0.037	0.039	0.042	0.038	0.042
Average Analytic SE	CC	0.045	0.042				
	FD	0.033	0.031				
	MR	0.040	0.036	0.038	0.040	0.037	0.040

Table 16 Results for missing continuous covariate Z in setting 6

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.891	0.822				
	FD	0.951	0.912				
	MR	0.946	0.911	0.948	0.948	0.948	0.948
Bias	CC	0.008	0.016				
	FD	-0.001	0.007				
	MR	-0.001	0.007	-0.001	-0.001	0.000	0.000
MC SE	CC	0.021	0.019				
	FD	0.016	0.015				
	MR	0.017	0.016	0.017	0.017	0.017	0.017
Average Analytic SE	CC	0.021	0.019				
	FD	0.016	0.015				
	MR	0.017	0.016	0.017	0.017	0.017	0.017

Table 17 Results for missing continuous covariate Z in setting 7

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.883	0.748				
	FD	0.957	0.873				
	MR	0.957	0.883	0.954	0.954	0.952	0.955
Bias	CC	0.020	0.038				
	FD	0.001	0.019				
	MR	0.001	0.019	0.001	0.001	0.001	0.001
MC SE	CC	0.032	0.031				
	FD	0.024	0.024				
	MR	0.025	0.024	0.024	0.025	0.024	0.025
Average Analytic SE	CC	0.031	0.030				
	FD	0.024	0.024				
	MR	0.025	0.024	0.025	0.025	0.024	0.025

Table 18 Results for missing continuous covariate Z in setting 8

Estimate		Scenario					
		1	2	3	4	5	6
Coverage Pr	CC	0.879	0.817				
	FD	0.944	0.872				
	MR	0.939	0.894	0.936	0.939	0.942	0.941
Bias	CC	0.018	0.031				
	FD	0.000	0.014				
	MR	0.000	0.014	0.000	0.000	0.001	0.001
MC SE	CC	0.041	0.038				
	FD	0.022	0.021				
	MR	0.024	0.023	0.024	0.024	0.023	0.024
Average Analytic SE	CC	0.039	0.037				
	FD	0.021	0.020				
	MR	0.023	0.022	0.023	0.024	0.022	0.023

For both missing binary and continuous covariate Z, we could find the MR estimate has lower coverage probability when $Pr(A = 1) = 0.8$ (setting 5) compared with other settings. Besides, $\hat{\psi}_{MR}$ has a lower coverage probability when A model was wrong. The range of marginal means of our simulation settings covered most practical settings where the multiply robust approach would have good performance.

2.4 Discussion

Based on the MR estimator $\hat{\psi}_{MR}$ for the average causal effect of a treatment in the presence of missing binary confounder data, we extent the method to the observational data with a missing continuous confounder. We demonstrated that $\hat{\psi}_{MR}$was robust to model misspecification that it was consistent for the average treatment effect if one of the four conditions held: model (i) for $E[Y|A, \mathbf{X}, Z]$ and model (iii) for $\Pr(R = 1|Y, A, \mathbf{X})$ were correct, model (ii) for $\Pr(A = 1|\mathbf{X}, Z)$ and model (iii) were correct, model (i) and model (iv) for $\Pr(Z = 1|R = 1, Y, A, \mathbf{X})$ were correct, and models (ii) and (iv) were correct. We also developed the asymptotic variance for $\hat{\psi}_{MR}$ when either missing binary or continuous variable exists. A simulation study was carried out to evaluated the estimate's performance in many aspects.

In this chapter, the MR estimator provides a promising unified approach to the estimation of causal effects in the presence of one missing binary/continuous confounder data in observational studies. In principal, the methodology can be extended to handle one or several categorical or continuous confounders **Z.** When all components of **Z** are discrete and have a small number of covariate levels, the conditional mean of any function $h(Y, A, \mathbf{X}, Z)$ given $(R = 1, Y, A, \mathbf{X})$ can be easily calculated as a sum over all

possible covariate levels of **Z.** When **Z** has at least one continuous component or a large number of possible covariate levels, the conditional mean function can be estimated by fitting a parametric working model for $E[h(Y, A, \mathbf{X}, Z)|R = 1, Y, A, \mathbf{X}]$ among the complete cases.

CHAPTER 3. AN INVERSE PROBABILITY WEIGHTING APPROACH FOR HISTORICAL CONTROL AUGMENTED RANDOMIZED CLINICAL TRIALS

3.1 Background

The randomized control trial (RCT) is often considered as the gold standard for evaluating the effectiveness or safety of various types of medical interventions. Patients in RCT are randomly allocated to different treatments under study so that patients across treatments are reasonably comparable in both known and unknown prognostic factors. After randomization, the two (or more) groups of patients are followed in exactly the same way, leaving the treatment that the patients receive as the main difference.

An alternative of RCT is the single arm trial where all patients in the current study will receive the new treatment, which the investigators believe to be superior. Nevertheless, due to the lack of randomization, the single arm trial may lead to biased estimates of the treatment effect when the comparability between the historical and current study is not ensured. Historical and current trial may differ in a number of ways including the change in supporting care, the distribution of patient characteristics, the method of the effect evaluation and the inter-institution variability. Thus the single arm trial is usually considered appropriate mostly in situations where the well-defined historical data are available or accrual is a problem in rare disease settings [17, 18].

A comprise between the single-arm trial and regular RCT for the new treatment is a RCT with the current control group augmented by relevant historical controls[19]. To avoid a systematic bias in such type of clinical trial design, the acceptability of the historical data should be carefully evaluated, in a prospective rather than a *post hoc* manner. Generally, the historical data should come from a randomized trial with the same

(or similar) patient eligibility and the same effect evaluation method. The historical trial should usually be performed in the same organization with largely the same investigators, and no apparent dissimilarity in other aspects of study design such as accrual rate [19]. Even under most stringent scrutiny, the prospectively selected historical controls are generally not considered as reliable as the randomized control. It is expected that some degree of bias might exist in historical data with both the magnitude and direction not predictable before the current trial.

Various approaches have been proposed for combining current and historical control to achieve the statistical efficiency while protecting against the potential biases [19-24]. Such methods include the test-then-pool[19], power priors[20] and hierarchical Bayesian models [21, 22](see [25] for an excellent review). For the situations that inconsistency of the two sources of data can be explained by the measured covariates, it is recognized that we could calibrate the current and historical controls via the covariate adjustment using either regression or propensity score methodology in Bayesian approach. [11, 26, 27] We could also consider the combining problem as a confounding problem in observational study. However, using inverse probability weighting (IPW) method for confounding problem hasn't been explored yet. Compared with Bayesian models, IPW method avoids the selection of prior and still help to improve the efficiency.

Our research is motivated by the need of an efficient trial design for finding novel treatments for amyotrophic lateral sclerosis (ALS), a devastating neurodegenerative disease currently with no cure, affecting about 30,000 Americans. Riluzole is the only FDA-approved treatment that has small but significant beneficial effect in prolonging life span and slowing declines in some motor function of ALS patients [28-30]. It is generally

reported that the median survival time of ALS patients from symptom onset is 2 to 5 years[31], but the survival time of individual ALS patients varies widely. There are many known prognostic factors that are associated with ALS patient survival times. Investigators found it important to include these known prognostic factors in the primary analysis of the clinical trials evaluating the survival effectiveness of Riluzole [30]. Similarly, it may be even more crucial to include known prognostic factors into any type of models for combining historical data in ALS trials. With a recent resurgence in the number of ALS clinical trials, more relevant historical information may be available for augmenting future ALS trials [32]. Specifically, a randomized, double-blind, placebo-controlled phase 3 trial on ALS (EMPOWER) was recently completed that contain 468 patients on the placebo group [33]. We are interested in developing an appropriate frequentist method to incorporate EMPOWER control data to potentially improve the efficiency of future trial design on ALS.

In the following sections, we first introduce our notation. Then we present the new IPW estimates for continuous and time-to-event response. We carry out a simulation study for the time-to-event response and compare the results from IPW estimates with other estimates (from not combining historical controls). Finally, we conclude our paper with a brief discussion.

3.2 Methods

In this section, we begin with the most straightforward outcome, continuous response, then we extend the similar idea to the time-to-event response, which is consistent with our motivation example.

3.2.1 IPW estimates for continuous

3.2.1.1 Notation

We collect data from one RCT and the qualified historical controls. There're two arms (with/without intervention) for three groups: treatment group in trial receiving intervention (group T), control group in trial not receiving intervention (group C) and historical control group not in trial not receiving intervention (group H). In Table 19, I_1 and I_2 are the indicators for three groups. We define the population B for the population not receiving treatment, from which we select group C and group H patients into the study.

Table 19 Indicators for three arms

Group / Indicator	T	C	H
I_1	1	0	0
I_2	0	1	0

Let $i = 1, \dots, n_T$ denote the subjects in group T; $j = 1, \dots, n_C$ denote the subjects in group C; $k = 1, \dots, n_H$ denote the subjects in group H; $l = 1, \dots, n_1 = n_C + n_H$ denote all the subjects in population B; $m = 1, \dots, n_2 = n_T + n_C$ denote all the subjects in RCT; $q = 1, \dots, N$ denote all the subjects in the study.

For each subject, Y denotes the observed response; $\mathbf{X}$ denotes the covariates; $e = \Pr(I_2 = 1|\mathbf{X})$ denotes the propensity score for population B. For subject l, $I_{2l} \sim Bernolli\left(e(\mathbf{X}_l, \beta) = \Pr\left(I_{2l} = 1|\mathbf{X}_l, \beta\right) = \frac{\exp(X_l'\beta)}{1+\exp(\mathbf{X}_l'\beta)}\right)$. The estimated parameter β and propensity score $e_l = e(X_l, \beta)$ could be obtained through logistic regression from population B. We define random variables (R.V.):

- $S = \frac{I_2}{e}Y$; $W = \frac{(1-I_2)}{1-e}Y$; $K = \frac{(1-I_2)}{e}Y$; $L = \frac{I_2}{1-e}Y$; $U = eY$.
- $S_l^* = S_l(\beta^*) = \frac{I_{2l}}{e_l(\beta^*)}Y_l$, where $\beta^*$is the truth of β.

  For the populations/groups, we have

- E_B^S denotes for the true mean of R.V. S from population B. $P_{n_1}(S_l) = \frac{1}{n_1}\sum_{l=1}^{n_1} S_l$ denotes for the sample mean of R.V. S using samples from population B.
- $Var_B^Y = E_B^{Y^2} - (E_B^Y)^2$ denotes for the true variance of R.V. Y from population B. $V_B^Y = P_{n_1}(Y_l^2) - \left[P_{n_1}(Y_l)\right]^2$ denotes the sample variance of R.V. Y using samples from population B.
- $Cov_C(Y,S) = E_C^{YS} - E_C^Y E_C^S$ denotes for the true covariance of R.V. Y and S from group C. Sample covariance $Q_C(Y,S) = P_{n_c}(Y_jS_j) - P_{n_c}(Y_j)P_{n_c}(S_j)$ denotes for the covariance of R.V. Y and S using samples from group C.

3.2.1.2 IPW Estimate for continuous response

Let θ denote for the treatment effects of the intervention, which is our parameter of interest

$$\theta = E(Y^1) - E(Y^0)$$

Y^1 and Y^0 denotes for the response from the subject receiving/not receiving intervention.

First, we could construct the estimate for θ only based on the trial data using the sample average:

$$\hat{\theta}_1 = u^1 - u^0 \tag{7}$$

where $u^1 = \overline{Y_T} = \frac{1}{n_T}\sum_{i\in T} Y_i$, $u^0 = \overline{Y_C} = \frac{1}{n_C}\sum_{j\in C} Y_j$. $Var(\hat{\theta}_1) = \frac{1}{n_T}Var_T^Y + \frac{1}{n_C}Var_C^Y$.

Next, we construct four estimates from combining current controls and historical controls.

1. Historical controls are combined into current controls and viewed as part of RCT on the arm without intervention. $\hat{\theta}_2$ is the estimates of the average treatment effects over the treated (ATT). $E(Y_B^1)$ means the expected response from population B when receive the treatment. Let $g = \Pr(I_1 = 1|X) = \frac{\Pr(I_1=1,X)}{\Pr(X)} = \frac{e}{1+e}$. $E(Y_B^1) = \sum_{q=1}^{N} \frac{(1-I_{1q})g_a}{1-g_q} Y_q$.

$$\hat{\theta}_2 = \frac{1}{\Pr(I_1 = 1)}[E(Y_T^1) - E(Y_B^1)] = \frac{1}{\frac{n_T}{N}} \times \frac{1}{N}\sum_{q=1}^{N}\left[I_{1q} - \frac{(1-I_{1q})g_q}{1-g_q}\right]Y_q \tag{8}$$

$$= \frac{1}{n_T}\left[\sum_{i=1}^{n_T} Y_i - \sum_{l=1}^{n_1} \frac{g_l}{1-g_l} Y_l\right] = \frac{1}{n_T}\left[\sum_{i=1}^{n_T} Y_i - \sum_{l=1}^{n_1} e_l Y_l\right] = u^1 - \frac{n_1}{n_T} P_{n_1}(U_l)$$

Thus we have

$$Var(\hat{\theta}_2) = \frac{1}{n_T} V_T^Y + \frac{n_1^2}{n_T^2} \cdot \frac{1}{n_1} Var(U) = \frac{1}{n_T} V_T^Y + \frac{n_1}{n_T^2} V_B^U \tag{9}$$

2. We could construct the unbiased estimates

$$\hat{\theta}_{r+2} = \hat{\theta}_1 + a\,\Delta_r, r = 1, \dots, 3. \tag{10}$$

Thus $Var(\hat{\theta}_{r+2}) = Var(\hat{\theta}_1 + a\,\Delta_r)$. Here Δ_r estimates the effect of getting involved in RCT in population B. To get Δ_r, we have three kinds of estimates of response from population B

$$d_1 = P_{n_1}(Y_l), d_2 = P_{n_1}(S_l), d_3 = P_{n_1}(W_l) \tag{11}$$

So we could construction

a. $\Delta_1 = d_2 - d_3 = P_{n_1}(S_l - W_l)$.

b. $\Delta_2 = d_1 - d_2 = P_{n_1}(Y_l - S_l) = \frac{1}{n_1}\left(\sum_{k\in H} Y_k - \sum_{j\in C}\frac{1-e_j}{e_j}Y_j\right)$.

c. $\Delta_3 = d_1 - d_3 = P_{n_1}(Y_l - W_l) = \frac{1}{n_1}\left(\sum_{j\in C} Y_j - \sum_{k\in H}\frac{e_k}{1-e_k}Y_k\right)$.

To find the most efficient $\hat{\theta}_{r+2}$, we minimize $Var(\hat{\theta}_{r+2})$ and find the parameter $a^* = -\frac{Cov(\hat{\theta}_1, \Delta_r)}{Var(\Delta_r)}$. We simply use $\hat{\theta}_{r+2}$ for $\hat{\theta}_{r+2}|_{a^*}$ and we have

$$Var(\hat{\theta}_{r+2}) = Var(\hat{\theta}_1) - \frac{Cov(\hat{\theta}_1, \Delta_r)^2}{Var(\Delta_r)} = (1-\rho^2)Var(\hat{\theta}_1) \le Var(\hat{\theta}_1) \tag{12}$$

Here $\rho = \frac{cov(\hat{\theta}_1, \Delta_r)}{\sqrt{Var(\hat{\theta}_1)Var(\Delta_r)}}$. The large ρ^2 is, the more efficient $\hat{\theta}_{r+2}$ is.

3.2.2 IPW estimate for time-to-event response

3.2.2.1 Notation

- I_1 and I_2 are the indicators for group T and group C.
- n_C, n_1 and n_2 dare the number of subjects in group C, group C + group H and RCT respectively;
- $Y_i = (\delta_i, T_i)$ denotes the response for subject i, where δ_i is the censoring indicator ($\delta_i = 0$ means subject i is censored) and T_i is the survival time.
- $\mathbf{X}_i$ is the covariates of subject i and $\pi_i = \Pr(I_{2i} = 1|\mathbf{X}_i)$ is the propensity score of being in group C.

3.2.2.2 Efficient score function

In a standard RCT with two arms, let I be the indicator for the treatment arm and n be the total number of subjects in the study. Based on cox model assuming proportional hazard assumption holds, we have the hazard rate function

$$\lambda(t) = \lambda_0(t)\exp(\theta I_i)$$

And the partial likelihood could be written as

$$L(\theta) = \prod_{i=1}^{n} \left(\frac{\exp(\theta I_i)}{\Sigma_{j \in R_i}\left(\exp\left(\theta I_j\right)\right)} \right)^{\delta_i}$$

where $R_i = \{j: T_j \geq T_i\}$. The log partial likelihood:

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \delta_i \{\theta I_i - \log[\Sigma_{j \in R_i}\left(\exp\left(\theta I_j\right)\right)]\}$$

Therefore we have the score function:

$$\frac{\partial l}{\partial \theta} = \sum_{i=1}^{n} \delta_i \left\{ I_i - \frac{\Sigma_{j \in R_i}\left(I_j \exp\left(\theta I_j\right)\right)}{\Sigma_{j \in R_i}\left(\exp\left(\theta I_j\right)\right)} \right\} = 0$$

We use MLE to get $\hat{\theta}$. Let $\tilde{l}_\theta(Y_i) = \delta_i \left\{ I_i - \frac{\Sigma_{j \in R_i}\left(I_j \exp\left(\theta I_j\right)\right)}{\Sigma_{j \in R_i}\left(\exp\left(\theta I_j\right)\right)} \right\}$. Then $\frac{\partial l}{\partial \theta} = \sum \tilde{l}_\theta(Y_i)$.

According to Eq(31) from [34], the efficient score function $l^*_{\theta,\Lambda}$ for subject i is

$l^*_{\theta,\Lambda}(Y_i) = \delta_i \left\{ I_i - \frac{\Sigma_{j \in R_i}\left(I_j \exp\left(\theta I_j\right)\right)}{\Sigma_{j \in R_i}\left(\exp\left(\theta I_j\right)\right)} \right\} - \exp(\theta I_i) \int_0^{t_i} \left\{ I_i - \frac{\Sigma_{j \in R_k}\left(I_j \exp\left(\theta I_j\right)\right)}{\Sigma_{j \in R_k}\left(\exp\left(\theta I_j\right)\right)} \right\} d\Lambda(t_k)$After

plugging the estimated baseline functiion $d\hat{\Lambda}(t_k) = \hat{h}(t_k) = \delta_k \frac{1}{\Sigma_{j \in R_k}\left(\exp\left(\theta I_j\right)\right)}$, we have

$$\hat{l}^*_{\theta,\hat{\Lambda}}(Y_i) = \delta_i \left\{ I_i - \frac{\Sigma_{j \in R_i}\left(I_j \exp\left(\theta I_j\right)\right)}{\Sigma_{j \in R_i}\left(\exp\left(\theta I_j\right)\right)} \right\}$$

$$- \exp(\theta I_i) \sum_{t_k \leq t_i} \left\{ \left[I_i - \frac{\Sigma_{j \in R_k}\left(I_j \exp\left(\theta I_j\right)\right)}{\Sigma_{j \in R_k}\left(\exp\left(\theta I_j\right)\right)} \right] \times \left[\delta_k \frac{1}{\Sigma_{j \in R_k}\left(\exp\left(\theta I_j\right)\right)} \right] \right\}$$

$$= \delta_i \left\{ I_i - \frac{\Sigma_{j \in R_i}\left(I_j \exp\left(\theta I_j\right)\right)}{\Sigma_{j \in R_i}\left(\exp\left(\theta I_j\right)\right)} \right\}$$

$$- \exp(\theta I_i) \sum_{t_k \leq t_i} \left\{ \delta_k \left[\frac{I_i}{\Sigma_{j \in R_k}\left(\exp\left(\theta I_j\right)\right)} - \frac{\Sigma_{j \in R_k}\left(I_{1j} \exp\left(\theta I_j\right)\right)}{\left(\Sigma_{j \in R_k}\left(\exp\left(\theta I_j\right)\right)\right)^2} \right] \right\}$$

This is consistent with Eq(2) from [35].

For inverse probability weighting (IPW) Cox model, $d\widehat{\Lambda}(t_k) = \hat{h}(t_k) = \frac{\delta_k}{\pi_k \times \Sigma_{j \in R_k}\left(\exp\frac{(\theta I_j)}{\pi_j}\right)}$ and the efficient score function is:

$$\hat{l}^*_{\theta,\widehat{\Lambda}}(Y_i) = \delta_i \left\{ I_i - \frac{\Sigma_{j \in R_i}\left(\frac{I_j \exp(\theta I_j)}{\pi_j}\right)}{\Sigma_{j \in R_i}\left(\frac{\exp(\theta I_j)}{\pi_j}\right)} \right\}$$

$$- \exp(\theta I_i) \sum_{t_k \le t_i} \left\{ \left[I_i - \frac{\Sigma_{j \in R_k}\left(\frac{I_j \exp(\theta I_j)}{\pi_j}\right)}{\Sigma_{j \in R_k}\left(\frac{\exp(\theta I_j)}{\pi_j}\right)} \right] \times \left[\delta_k \frac{1}{\pi_k \times \Sigma_{j \in R_k}\left(\frac{\exp(\theta I_j)}{\pi_j}\right)} \right] \right\}$$

$$= \delta_i \left\{ I_i - \frac{\Sigma_{j \in R_i}\left(\frac{I_j \exp(\theta I_j)}{\pi_j}\right)}{\Sigma_{j \in R_i}\left(\frac{\exp(\theta I_j)}{\pi_j}\right)} \right\}$$

$$- \exp(\theta I_i) \sum_{t_k \le t_i} \left\{ \frac{\delta_k}{\pi_k} \left[\frac{I_i}{\Sigma_{j \in R_k}\left(\frac{\exp(\theta I_j)}{\pi_j}\right)} - \frac{\Sigma_{j \in R_k}\left(\frac{I_j \exp(\theta I_j)}{\pi_j}\right)}{\left(\Sigma_{j \in R_k}\left(\frac{\exp(\theta I_j)}{\pi_j}\right)\right)^2} \right] \right\}$$

Here the baseline function is estimated as $d\widehat{\Lambda}(t_k) = \hat{h}(t_k) = \delta_k \frac{1}{\Sigma_{j \in R_k}\left(\frac{\exp(\theta I_j)}{\pi_j}\right)}$ according to Eq(35) from [34]. θ_0 denotes the truth of θ. The information matrix based on the score function is

$$\tilde{I}_{\theta_0} = -\frac{1}{n_2} \times \frac{\partial l^2}{\partial \theta^2}|_{\theta_0} =$$

$$-\frac{1}{n_2}\sum_{i=1}^{n_2} -\delta_i \frac{\Sigma_{j\in R_i}\left[\exp(\theta_0 I_j) I_j^2\right]\Sigma_{j\in R_i}(\exp(\theta_0 I_j)) - \Sigma_{j\in R_i}[I_{1j}\exp(\theta_0 I_j)]\Sigma_{j\in R_i}[I_j \exp(\theta_0 I_j)]}{\left[\Sigma_{j\in R_i}(\exp(\theta_0 I_j))\right]^2}$$

$$= \frac{1}{n_2}\sum_{i=1}^{n} \delta_i \left[\frac{\Sigma_{j\in R_i}\left[\exp(\theta_0 I_j)\, I_j^2\right]}{\Sigma_{j\in R_i}(\exp(\theta_0 I_j))} - \left(\frac{\Sigma_{j\in R_i}\left[I_j \exp(\theta_0 I_j)\right]}{\Sigma_{j\in R_i}(\exp(\theta_0 I_j))}\right)^2\right]$$

Therefore we have

$$Var(\theta) = \left(n * \tilde{I}_{\theta_0}^{-1}\right)\sum_{i=1}^{n}\left(\frac{\hat{l}^*_{\theta,\hat{\Lambda}}(Y_i)}{\pi_i}\right)^2 \left(n * \tilde{I}_{\theta_0}^{-1}\right)$$

According to Eq(32) from [34], the Information matrix is

$$\tilde{I}_{\theta_0} = P_0\left(\exp(\theta_0 I_i)\int_0^{\tau}\left\{I_i - \frac{\Sigma_{j\in R_k}\left(I_j \exp(\theta_0 I_j)\right)}{\Sigma_{j\in R_k}(\exp(\theta_0 I_j))}\right\}^{\otimes 2} \Pr(T \geq t_k | I_i)\, d\Lambda_0(t_k)\right)$$

Based on (3.21) from Fleming and Harrington [36], we have

$$V(\theta_0, t_i) = \frac{\Sigma_{j\in R_i}\left[\exp(\theta_0 I_j)\, I_j^2\right]}{\Sigma_{j\in R_i}(\exp(\theta_0 I_j))} - \left(\frac{\Sigma_{j\in R_i}\left[I_j \exp(\theta_0 I_j)\right]}{\Sigma_{j\in R_i}(\exp(\theta_0 I_j))}\right)^2$$

It's been proved that (3.23) is equivalent to (3.21) and in our setting

$$V(\theta_0, t_i) = \frac{1}{n_2}\cdot\frac{\sum_{k=1}^{n}\left[\left(I_i - \frac{\Sigma_{j\in R_k}\left[I_j \exp(\theta_0 I_j)\right]}{\Sigma_{j\in R_k}(\exp(\theta_0 I_j))}\right)^{\otimes 2} Y_k(t_i)\exp(\theta_0 I_i)\right]}{\Sigma_{j\in R_i}(\exp(\theta_0 I_j))}$$

$$= \frac{\exp(\theta_0 I_i)\sum_{k=1}^{n}\left[\left(I_i - \frac{\Sigma_{j\in R_k}\left[I_j \exp(\theta_0 I_j)\right]}{\Sigma_{j\in R_k}(\exp(\theta_0 I_j))}\right)^{\otimes 2} \frac{Y_k(t_i)}{n_2}\right]}{\Sigma_{j\in R_i}(\exp(\theta_0 I_j))}$$

$$= \frac{\exp(\theta_0 I_i)\sum_{k=1}^{n}\left[\left(I_{1i} - \frac{\Sigma_{j\in R_k}\left[I_j \exp(\theta_0 I_j)\right]}{\Sigma_{j\in R_k}(\exp(\theta_0 I_j))}\right)^{\otimes 2} \Pr(T \geq t_k | I_i)\right]}{\Sigma_{j\in R_i}(\exp(\theta_0 I_j))}$$

After plugging the estimated baseline functiion $d\widehat{\Lambda}_0(t_k) = \hat{h}_0(t_k) = \delta_k \frac{1}{\Sigma_{j\in R_k}\left(\frac{\exp(\theta I_j)}{\pi_j}\right)}$,

we could prove that two information matrix formulas are the same.

3.2.2.3 IPW estimate for time-to-event response

Parameter of interest θ is the log hazard ratio between treated and control. We first construct the estimate for θ only based on the trial data using Cox model. θ_0 denote the truth of θ. Using Taylor expansion, we have

$$\tilde{I}_{\theta_0}\sqrt{n_2}\left(\hat{\theta}_1 - \theta_0\right) = \frac{1}{\sqrt{n_2}}\sum_{i=1}^{n_2} \hat{l}^*_{\theta_0,\hat{\Lambda}}(Y_i) + O_p(1)$$

$$\hat{\theta}_1 = \frac{1}{n_2}\sum_{i=1}^{n_2}\left[\theta_0 + \tilde{I}_{\theta_0}^{-1}\hat{l}^*_{\theta_0,\hat{\Lambda}}(Y_i)\right] + \frac{1}{\sqrt{n_2}}O_p(1)$$

Here $\tilde{I}_{\theta_0}^{-1} = \frac{1}{n_2}\sum_{i=1}^{n_2}\delta_i\left[\frac{\Sigma_{j\in R_i}\left[\exp(\theta_0 I_{1j})I_j^2\right]}{\Sigma_{j\in R_i}(\exp(\theta_0 I_{1j}))} - \left(\frac{\Sigma_{j\in R_i}[I_{1j}\exp(\theta_0 I_{1j})]}{\Sigma_{j\in R_i}(\exp(\theta_0 I_{1j}))}\right)^2\right]$ is the inverse of information matrix.

$$\hat{l}^*_{\theta_0,\hat{\Lambda}}(Y_i) = \delta_i\left\{I_{1i} - \frac{\Sigma_{j\in R_i}\left(I_{1j}\, exp\left(\theta_0 I_{1j}\right)\right)}{\Sigma_{j\in R_i}\left(exp\left(\theta_0 I_{1j}\right)\right)}\right\}$$

$$- exp(\theta_0 I_{1i})\sum_{t_k\le t_i}\left\{\delta_k\left[\frac{I_{1i}}{\Sigma_{j\in R_k}\left(exp\left(\theta_0 I_{1j}\right)\right)} - \frac{\Sigma_{j\in R_k}\left(I_{1j}\, exp\left(\theta_0 I_{1j}\right)\right)}{\left(\Sigma_{j\in R_k}\left(exp\left(\theta_0 I_{1j}\right)\right)\right)^2}\right]\right\}$$

is the efficient score function for subject i.

Meanwhile, we obtain the log hazard ratio $\hat{p}$ from historical data and control data of RCT using IPW Cox model.

$$\tilde{I}_0\sqrt{n_1}(\hat{p} - 0) = \tilde{I}_0\sqrt{n_1}\hat{p} = \frac{1}{\sqrt{n_1}}\sum_{i=1}^{n_1}\hat{l}^*_{\hat{p},\hat{\Lambda}_p}(Y_i) + O_p(1)$$

$$\hat{p} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\tilde{I}_0^{-1}\hat{l}^*_{\hat{p},\hat{\Lambda}_p}(Y_i)\right] + \frac{1}{\sqrt{n_1}}O_p(1)$$

The information matrix $\tilde{I}_0 = \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{1}{\pi_i}\delta_i\left[\frac{\Sigma_{j\in R_i}\frac{I_{1j}^2}{\pi_j}}{\Sigma_{j\in R_i}\frac{1}{\pi_j}} - \left(\frac{\Sigma_{j\in R_i}\frac{I_{1j}}{\pi_j}}{\Sigma_{j\in R_i}\frac{1}{\pi_j}}\right)^2\right]$. $\pi_j = \Pr\left(I_{2j} = 1|\mathbf{X}_j\right)$ is the propensity score for being in group C. $\hat{\Lambda}_p$ is the estimated baseline hazard function from IPW Cox model.

Then we use the linear combination of $\hat{\theta}_1$ and $\hat{p}$ to construct the unbiased estimate $\hat{\theta}_2 = \hat{\theta}_1 + a\cdot\hat{p}$. To minimize the variance of $\hat{\theta}_2$, we have

$$Var\left(\hat{\theta}_2\right) = Var\left(\hat{\theta}_1 + a\hat{p}\right)$$

$$= Var\left(\hat{\theta}_1\right) + 2aCov\left(\hat{\theta}_1,\hat{p}\right) + a^2Var(\hat{p})$$

Let $a^* = -\frac{Cov\left(\hat{\theta}_1,\hat{p}\right)}{Var(\hat{p})}$. We simply use $\hat{\theta}_2$ for $\hat{\theta}_2|_{a^*}$. Then we have

$$Var\left(\hat{\theta}_2\right) = Var\left(\hat{\theta}_1\right) - \frac{Cov\left(\hat{\theta}_1,\hat{p}\right)^2}{Var(\hat{p})} = (1-\rho^2)Var\left(\hat{\theta}_1\right) \le Var\left(\hat{\theta}_1\right) \tag{13}$$

Here $\rho = \frac{cov\left(\hat{\theta}_1,\hat{p}\right)}{\sqrt{Var\left(\hat{\theta}_1\right)Var(\hat{p})}}$. The larger ρ^2 is, the more efficient $\hat{\theta}_2$ is.

$$Cov\left(\hat{\theta}_1,\hat{p}\right) = Cov\left(\frac{1}{n_2}\sum_{i=1}^{n_2}\left[\theta_0 + \tilde{I}_{\theta_0}^{-1}\hat{l}^*_{\theta_0,\hat{\Lambda}}(Y_i)\right], \frac{1}{n_1}\sum_{i=1}^{n_1}\left[\frac{1}{\pi} * \tilde{I}_0^{-1}\hat{l}^*_{\hat{p},\hat{\Lambda}_p}(Y_i)\right]\right)$$

$$= Cov\left(\frac{1}{n_2}\sum_{m=1}^{n_T+n_C}\left[\tilde{I}_{\theta_0}^{-1}\hat{l}^*_{\theta_0,\hat{\Lambda}}(Y_m)\right], \frac{1}{n_1}\sum_{l=1}^{n_C+n_H}\left[\frac{1}{\pi} * \tilde{I}_0^{-1}\hat{l}^*_{\hat{p},\hat{\Lambda}_p}(Y_l)\right]\right)$$

$$= \frac{1}{n_1n_2}\sum_{m=1}^{n_T+n_C}\sum_{l=1}^{n_C+n_H} Cov\left(\tilde{I}_{\theta_0}^{-1}\hat{l}^*_{\theta_0,\hat{\Lambda}}(Y_m), \frac{1}{\pi} * \tilde{I}_0^{-1}\hat{l}^*_{\hat{p},\hat{\Lambda}_p}(Y_l)\right)$$

$$= \frac{\tilde{I}_{\theta_0}^{-1}}{n_1 n_2}\left[\sum_{j=1}^{n_C}\sum_{l=1}^{n_C} Cov\left(\hat{l}^*_{\theta_0,\hat{\Lambda}}(Y_j), \frac{1}{\pi} * \hat{l}^*_{\hat{p},\hat{\Lambda}_p}(Y_l)\right)\right]\tilde{I}_0^{-1}$$

$$= \frac{n_C}{n_1 n_2}\tilde{I}_{\theta_0}^{-1}\left[Cov\left(\hat{l}^*_{\theta_0,\hat{\Lambda}}(Y^C), \frac{1}{\pi} * \hat{l}^*_{\hat{p},\hat{\Lambda}_p}(Y^C)\right)\right]\tilde{I}_0^{-1} \tag{14}$$

Y^C is the response of subjects from group C. Also we can get

$$Var(\hat{p}) = \frac{1}{n_1}\tilde{I}_0^{-1} Var\left[\hat{l}^*_{\hat{p},\hat{\Lambda}_p}(Y^H)\right]\tilde{I}_0$$

Here Y^H is response of the subject from group H. Based on (14), we could calculate a^* and $\hat{\theta}_2$ with more efficiency gain.

3.3 Simulation

We carried out a simulation study to evaluation the performance of IPW estimate $\hat{\theta}_2$ especially the efficiency gain in dealing with time-to-event response. With the following steps, we simulated the dataset.

1. Simulate covariate $\mathbf{X}$

   $X_1 \sim \text{Beta}(\alpha, \beta)$, $(X_2, X_3, X_4) \sim N((0,0,0)^T, \mathbf{I_3})$.

2. Simulate the propensity score (PS)

   $e = \Pr(I_1 \; or \; I_2 = 1) = X_1 \sim \text{Beta}(\alpha, \beta)$. e denotes the probability of being in RCT.

3. Simulate the group indicator A

   $A \sim Multinomial\left(p = \left(\frac{2}{3}e, \frac{e}{3}, 1-e\right)\right)$. In RCT, we set $\frac{n_T}{n_C} = 2$ to mimic the proportion of patients in ALS study. PS for group T, C and H is $\frac{2}{3}e$, $\frac{1}{3}e$ and $1-e$

respectively (see Table 20). Therefore, in group T, $e \sim \text{Beta}(\alpha+1, \beta)$; in group C, $e \sim \text{Beta}(\alpha+1, \beta)$; in group H, $e \sim \text{Beta}(\alpha, \beta+1)$.

Table 20 Multinomial distribution of group indicator A

Group	T	C	H
A	1	2	3
PS	$\frac{2}{3}e$	$\frac{1}{3}e$	$1-e$

4. Simulate the survival time S.

In group C, let $S \sim \exp(\frac{1}{\beta_C})$, where $\beta_C = \frac{ln2}{T_{med}^C}$. $T_{med}^C = 20.5$ is the median survival time of group C. In group T: $S \sim \exp(\frac{1}{\beta_T})$, where $\beta_T = \frac{\beta_C}{HR}$. Hazard ratio $HR = 0.7$ for group T v.s. C in RCT. Then we used Gaussian copula (parameter 0.3) to get the joint distribution of (S, e) based on distribution of S and e in group C and T respectively. Therefore we obtained the conditional distribution $(S|e)$. With PS e generated in previous step, we generated S. In group H we generated S using the same conditional distribution $(S|e)$ as group C, since their distribution should be the same given the same PS.

5. Simulate censoring time C

$C \sim Uniform[0,12] + 24$. We mimic the ALS study in which patients are uniformly enrolled within 12 months and followed 24 month at most.

6. Survival time T with right censoring $T = \min(S, C)$. $\delta = 1$ indicate an event.

We have 2 scenarios with different parameters of Beta distribution: $\alpha = 4, \beta = 16$; $\alpha = 12, \beta = 8$. When $\alpha = 4, \beta = 16$, we mimic the real situation with more historical data involved. The proportions of # of subjects in group T/C/H are 13.3%, 6.7%

and 80.0%, respectively. When $\alpha = 12, \beta = 8$, we had less historical data available. The proportions of # of subjects in T/C/H are 40%, 20% and 40%, respectively. In each scenario, we simulated 1000 MC datasets, each with 1000 subjects.

3.4 Results

For the estimation process, we used logistic regression (assuming that model is correctly specified) to obtain the estimated the PS $\hat{e}$. Then $\hat{e}$ was used for the estimation of $\hat{\theta}_2$. To evaluate the impact of the logistic regression model to our approach, we created another estimate $\tilde{\theta}_2$ which use the true PS during the whole estimation. The simulation results are as following.

Table 21 Results for scenario 1

	Mean	Bias	SD	MC SD	MSE	Coverage Prob
$\hat{\theta}_1$	-0.3657	-0.0091	0.1895	0.1888	0.0357	95.50%
$\hat{\theta}_2$	-0.3644	-0.0077	0.1454	0.1386	0.0193	96.00%
$\tilde{\theta}_2$	-0.3651	-0.0084		0.1394	0.0195	

Table 22 Results for scenario 2

	Mean	Bias	SD	MC SD	MSE	Coverage Prob
$\hat{\theta}_1$	-0.3503	0.0064	0.1090	0.1036	0.0108	96.10%
$\hat{\theta}_2$	-0.3546	0.0021	0.0857	0.0821	0.0067	95.90%
$\tilde{\theta}_2$	-0.3548	0.0018		0.0839	0.0070	

The truth of $\theta = \ln(HR) = \ln(0.7) = -0.3567$. In both scenarios (see Table 21 and Table 22), the MSE of $\hat{\theta}_2$ is much smaller than that of $\hat{\theta}_1$ and the coverage probability of $\hat{\theta}_2$ is close to the nominal level. The performance of $\tilde{\theta}_2$ is very similar as

$\hat{\theta}_2$, which shows the logistic regression model for PS e has little impact on the IPW estimate $\hat{\theta}_2$.

3.5 Discussion

Our simulation results show that $\hat{\theta}_2$ is an efficient estimate of the treatment effects from RCT. The information borrowed from historical data could greatly improve the precision of the estimation. We get a smaller SD and MSE by utilizing the historical control data in the approach. The coverage probability is close to nominal level.

When applying the IPW estimates, we might think these two ways. 1). Before the RCT with the chosen historical control data, we could pre-design the sample size of RCT. Since $Var(\hat{\theta}_2) = (1-\rho^2)Var(\hat{\theta}_1)$, with some assumption of $\rho = \frac{cov(\hat{\theta}_1, \hat{h})}{\sqrt{Var(\hat{\theta}_1)Var(\hat{h})}}$, we could have an appropriate estimated variance of the log hazard ratio (effect size), which is often critical to the sample size calculation. 2). After the RCT, we could use the IPW approach to analyze the RCT data together with historical control data to obtain an unbiased, more efficient estimate for the treatment effect from RCT.

CHAPTER 4. THE ASSOCIATIONS OF DRUGS WITH ACUTE MYOCARDIAL INFARCTION: BIAS CORRECTION, GLOBAL PROFILING AND INFERENCE ON INDIVIDUAL DRUG

4.1 Background

In typical non-randomized studies to investigate drug-outcome associations, risk measures are calculated based on models adjusting for an extended list of possible confounders. Many such studies are conducted in electronic health records databases (EHR) because of the wide adoption and availability of EHRs, and their population-wide and longitudinal coverage of patients. It has been recognized that EHRs are created for routine clinical care and administrative purposes but not for research, and thus they may harbor database biases due to, for example, failure or incomplete capture of exposure and covariates, hindering accurate estimations of causal effects. More likely than not, such database biases exist, hence the conventional estimate of the strength of a drug-outcome association, for example, from regressions, or even from more sophisticated propensity-score based methods, may still be biased, leading to attenuated or exaggerated association assessment. In addition, bias could be introduced due to model mis-specifications and compound the distortion of the estimate of the drug-outcome association. Either error causes public health concerns: Missing a harmful drug-outcome association can place patients at risk of adverse events, and exaggeration of drug-outcome association can deter patients from taking beneficial drugs. For example, previous epidemiologic studies suggested that estrogen plus progestin therapy prevented coronary heart disease, but a definitive, prospective randomized trial, [37], found harm of hormone therapy (hazard ratio 1.29 over placebo). The study stopped prematurely because an excessive risk of

invasive breast cancer was found associated with hormone replacement therapy, contrary to findings of no association from previous epidemiologic studies [38].

The database/model bias (henceforth the bias) could be characterized by a collection of drug-outcome association estimates from drugs which are known to be unrelated to the outcome events. For instance, the empirical distribution (Efron [39]) of risk estimates of these drugs can serve an characterization of the bias beyond what has been controlled in the analysis. The mean of such a distribution captures the bias due to limitations of the database or analytical methods. The idea of empirical nulls has been used in the analysis of EHR data (Schuemie, Ryan [40]). However, the previous focus in those applications has been on hypothesis testing rather than estimation. In a large EHR, accurate estimation of risk is much more prominent and relevant than hypothesis testing as the very magnitude of the strength is critical for decision-making. In addition, the characterization of estimation uncertainty by conventional frequentist method (e.g. confidence intervals) suffers from an unnatural interpretation (Diamond and Kaul [41]). Bayes inference offers an alternative method with heuristically appealing interpretation on estimation uncertainty using posterior distributions, particularly when the prior is a tangible and physically well-defined population (e.g. risk measures of all drugs for a given outcome). Along this line, empirical Bayes approach offers a reasonable solution (Robbins [42]), where the prior distribution is estimated from the data. The empirical Bayes approach has two extra advantages. First, there is a gain in accuracy for point estimate through borrowing information from other units (Efron [39]). Second, the prior distribution itself offers important knowledge on the state of the global drug-outcome associations.

In this paper, we report an analysis strategy that addresses two issues in pharmaco-epidemiology studies: bias reduction and precision improvement. We will use the empirical null distribution to correct bias and a new Bayes deconvolution method for an empirical Bayes inference. We focus our interest on a specific outcome, acute myocardial infarction (AMI) in a standardized subset of state-wise database. Our method represents a general strategy for future studies of drug-outcome association for a wide range of diseases.

In the following sections, we will first describe components of our method including data source, a self-controlled case series method for the estimation of incidence risk ratio of a given drug, a bias calibration procedure and a Bayes deconvolution method to estimate the prior distribution of the true associations for all drugs. The method is then applied to studying drug association with AMI. Furthermore, we conduct a simulation study to evaluate the performance of the method. Finally, we conclude the article with discussions.

4.2 Methods

4.2.1 Data source

The study is conducted using existing data, the Regenstrief Institute's electronic health records from 1/1/2004 to 12/31/2009 in the common data model,[43] which contains de-identified healthcare data of approximately 2.2 million unique patients from INPC. The common data model is a representative, longitudinal sample of the Indiana population in terms of prescriptions, claims, diagnoses and labs. We restrict our analysis

to patients of 40 years and older because of data use agreement with a certain insurance payer.

In the calculation of the risk ratio of AMI for each drug, the exposure period and the events of interest are defined as follows. We use drug eras in the "drug era" table in the common data model to define the exposure to a drug. A drug era consists of drug prescriptions, dispenses or refill claims that are recorded in successive periods within 30 days of each other and are combined to form one continuous period.[44] We added 90 days to the end of a drug era for the capture of AMI events for its possible lingering risk effect. The drug exposure is the collection of the drug eras. AMI is defined as the occurrence of at least one narrow diagnostic code ICD-9-CM 410* "Acute myocardial infarction" with either at least one diagnostic procedure code within 30 days prior to diagnostic code or at least one therapeutic procedure code within 60 days after the diagnostic code.

4.2.2 Self-controlled case series method

We used a self-controlled case series method (SCCS[45]) to estimate the incidence risk ratios with respect to AMI for all drugs with a sample size at least 10. SCCS is an intuitive method in controlling any known or unknown confounding as it uses each case (patient with the event under investigation) as his or her own control. It can be used to assess the association between an acute event and a transient exposure using cases only, without the need of a control group[45, 46]. The method estimates the ratio of the rate of events (number of events per unit time) in the exposure periods of a drug and the rate of events in the non-exposure periods during the entire observation period for a patient. A less than 1 rate ratio suggests that the drug has protective effect (or negative

association), whereas a rate ratio greater than 1 suggests a harmful effect (or positive association). If there were a positive association between a drug and a condition, one would expect higher event rates for the on-drug eras.

SCCS is appealing in database studies because of two major advantages.[46] First, it uses cases only and therefore is economical in terms of both time and cost. Second, it elegantly controls for both measured and unmeasured time-invariant confounders by using each subject as his own control, without explicitly modeling and adjusting confounders.

4.2.3 Bias calibration

Let RR denote the true incidence risk ratio of AMI for a drug, and $\widehat{RR}$ denote the estimated risk ratio. Correspondingly, $\theta = \ln RR$ and $\hat{\theta} = \ln \widehat{RR}$ are the logarithms of the true and the estimated risk ratios respectively. Applying SCCS to the INPC data, we obtained the point estimates of $\hat{\theta}$ and its standard error σ for each drug.

The bias for the estimation of the association of a drug with AMI Δ is defined as $E(Z) - \frac{\theta}{\sigma}$, where $Z = \frac{\hat{\theta}}{\sigma}$. Although we cannot identify the bias for each individual drug, it is possible to estimate the average of Δ from a set of drugs that are known to have no associations with AMI. Such an averaged bias can be used to correct risk ratio estimate to reduce bias. OMOP had identified 66 drugs as having no association with AMI (negative controls; see the table in Appendix for drug names), using information from the Food and Drug Administration (FDA) drug labels, literature, and randomized clinical trials to establish the drug-outcome associations and non-associations (Ryan, Schuemie [47]). The Z-statistics from these negative controls ideally would follow a standard normal distribution if there was no bias. The mean of Z-statistics of the negative controls $\bar{Z}_0$ in

general offers some evidence towards the existence of bias and we use it for the estimation of the average (across drugs) of Δ.

4.2.4 Bayes deconvolution method

Let $p(.)$ and $\hat{p}(.)$ denote the probability density function of RR and its estimate respectively. Our objectives were to compute (a) $\hat{p}(RR)$ that could then be used to estimate the proportions of drugs in the database with various strengths of association with AMI, and to derive (b) $\hat{p}\left(RR|Z^*, \sigma^*, \hat{\Delta}^*\right)$, the posterior distribution of RR for a new drug with summary data (Z^*, σ^*) and the estimated bias $\hat{\Delta}^*$, needed in making inference on the association of the new drug with AMI.

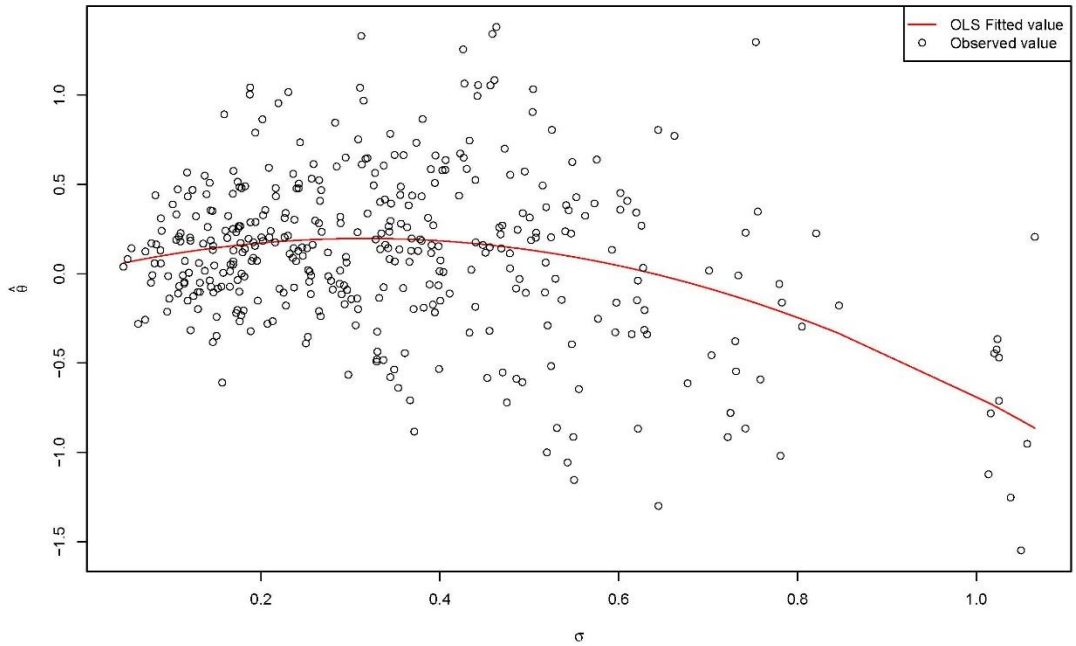
We considered the following model

$$Z = \frac{\hat{\theta}}{\sigma} = \frac{\theta}{\sigma} + \epsilon + \Delta = \frac{f(\sigma) + \delta}{\sigma} + \epsilon + \Delta, \epsilon \sim N(0,1) \qquad (15)$$

where we omit subscript (for drug) for simplicity of presentation. The main model assumption is that *RR* on the logarithm scale (θ) can be expressed as some function *f* of σ plus the random effect δ that has a marginal distribution $g(\delta)$. Further we assumed that δ and σ are independent. Since ϵ represents error due to sampling subjects, it is independent of (σ, δ). A scatter plot of $\hat{\theta}$ and σ (see Figure 1) exhibits a quadratic relationship between θ and σ, suggesting a model below:

$$f(\sigma) = \beta_1\sigma + \beta_2\sigma^2. \qquad (16)$$

Figure 1 Scatter plot and ordinary least square fit of $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\sigma}$



Note that the intercept for $f(\sigma)$ is absorbed in the distribution of δ. We now have

$$Z = \frac{\delta + \beta_1\sigma + \beta_2\sigma^2}{\sigma} + \epsilon + \Delta, \epsilon \sim N(0,1) \tag{17}$$

We then applied Efron's Bayes deconvolution method (Efron [48]). Specifically, we treated Δ as fixed and plugged in $\bar{Z}_0$ estimated from the negative controls. We assumed that the probability density function $g(\delta)$ belongs to a class of natural cubic splines. This class is fairly rich and covers or approximates most of the distributions in practice. It strikes a balance between nonparametric methods with strong robustness but slow convergence rate and parametric methods with root-N convergence rate but poor robustness. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) were used to select the degree of freedom of the natural cubic spline. To estimate $\hat{g}(\delta)$, we applied the maximum likelihood estimation (MLE) to summary data (Z_i, σ_i)

$i = 1, 2, \ldots, N$ as described by Efron [48]. The estimated $\hat{g}(\delta)$ was then used to compute $\hat{p}(RR)$ and $\hat{p}\left(RR|Z^*, \sigma^*, \hat{\Delta}^*\right)$ (Appendix for details).

4.3 Results

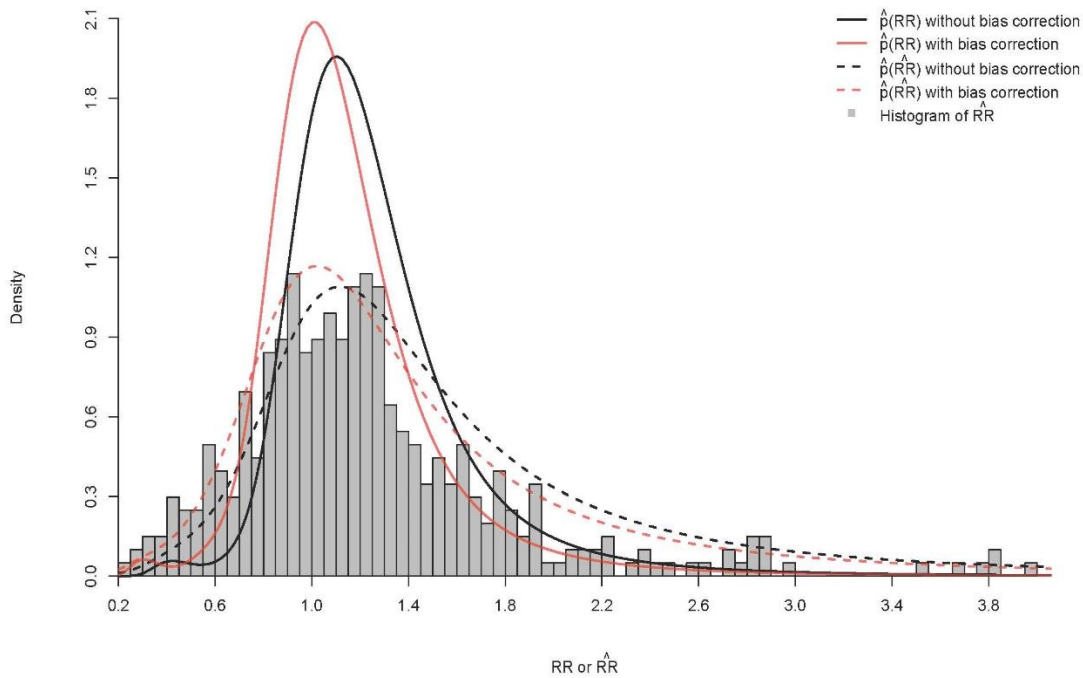
There are 465 drugs in the database suitable for SCCS analysis. Out of the 465 drugs, 59 drugs were excluded through a literature review as they are either used to prevent or to treat AMI. Two more drugs were excluded due to numerical instability. We kept the remaining 404 drugs for further analysis. The median and the interquartile range of the sample size of these drugs are 89 and 218 respectively.

We calculated the Z-statistics Z_i and standard error estimate $\sigma_i, i = 1, 2, \ldots, 404$. Among 66 true AMI negatives drugs (Ryan, Schuemie [47]), there are 39 drugs evaluable in the database. The mean of the Z-statistics of these drugs $\bar{Z}_0 = 0.31$, which is an estimate of the mean of the bias Δ in the drug-AMI associations. A test of the null-bias hypothesis results in a p-value of 0.053, indicating some statistical evidence on the existence of bias. We carried out all remaining analysis with and without the bias correction. For the analysis with the bias correction, we corrected Z-statistics by subtracting Ω from the original Z for all the drugs. Note that this procedure implies a constant bias for each drug, which may not hold in reality. Yet, drug-specific bias is not directly identifiable. In our simulation studies, we investigated the robustness of this procedure under the scenarios where the bias is not constant across drugs.

In Figure 2, we show the histogram of $\widehat{RR}$ for all 404 drugs, together with the estimated distribution of RR ($\hat{p}(RR)$) and $\widehat{RR}$ ($\hat{p}(\widehat{RR})$). It is clear that $\hat{p}(RR)$ after bias

correction shifts towards the left with respect to $\hat{p}(RR)$ without bias correction. The mean (SD) of RR is 1.17(0.34) (without bias correction) and 1.05(0.32) (with bias correction).

Figure 2 Histogram and the prior density of RR and $\widehat{\mathrm{RR}}$



Based on $\hat{p}(RR)$, 95% of RRs fall in [0.57,1.97] (without bias correction) and [0.43,1.79] (with bias correction). The probabilities of the included drugs having a risk ratio for AMI greater than 1, 1.5 and 2 (harmful effects) are 68.1%, 11.9%, 2.3% (without bias correction) and 51.5%, 7.0%, 1.4% (with bias correction). The probabilities of the included drugs having a risk ratio for AMI no greater than 1, 0.67 and 0.5 (null or protective effects) are 31.9%, 3.4% and 2.0% (without bias correction) and 48.5%, 6.4%, 3.1% (with bias correction). See Table 23 for more details.

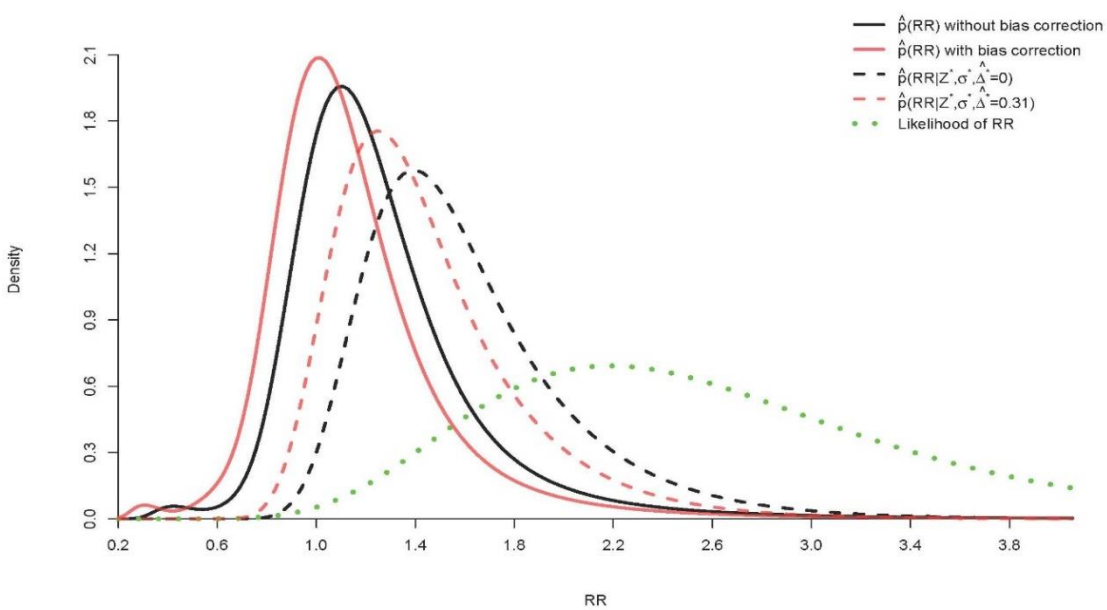
Table 23 Drug-outcome association inference from the estimated marginal distribution of RR ($\hat{p}(RR)$)

Bias	$\Pr(RR \leq c)$			$\Pr(RR > c)$		
	$c = 1$	$c = 0.67$	$c = 0.5$	$c = 1$	$c = 1.5$	$c = 2$
	Estimate (95%)	Estimate (95%)	Estimate (95%)	Estimate (95%)	Estimate (95%)	Estimate (95%)
0	0.319 [0.241,0.393]	0.034 [0.015,0.053]	0.020 [0.004,0.034]	0.681 [0.607,0.759]	0.119 [0.075,0.16]	0.023 [0.005,0.044]
0.31	0.485 [0.402,0.551]	0.064 [0.037,0.093]	0.031 [0.014,0.047]	0.515 [0.449,0.598]	0.070 [0.0,0.101]	0.014 [0.002,0.034]

Since $\hat{p}(RR)$ was estimated from data, we would like to understand the level of uncertainty in estimating $\hat{p}(RR)$ and its impact on the estimated posterior distribution. From 500 bootstrap samples, we obtained the 95% confidence intervals (CIs) of $\hat{p}(RR)$ (Table 23). Combining results with and without bias correction with consideration of uncertainty, as high as 4.4% of the drugs have an RR>2 (upper limit of the 95% confidence interval for $\Pr(RR > 2)$ without bias correction), or equivalently no more than 404*0.044=18 drugs are expected to have RR>2.

Suppose we have a hypothetical new drug with the same point estimate $\widehat{RR}^*$ and standard error estimates σ^* as drug "Aliskiren" ($\widehat{RR}^* = 2.19, \sigma^* = 0.345$ so that $Z^* = 2.27$), we computed the posterior distribution (dash lines in Figure 3).

Figure 3 Prior and posterior density of RR given $\widehat{\mathrm{RR}}^* = 2.19$, $\sigma^* = 0.345$



From the posterior distribution, the mean (SD) of RR is 1.52(0.37) (without bias correction) and 1.36(0.34) (with bias correction). 95% credible region of RR is [1.00,2.46] (without bias correction) and [0.91,2.18] (with bias correction). The

frequentist 95% CI for RR is $[1.11,3.74]$. So the precision gain is 45% (without bias correction) and 52% (with bias correction) with the understanding that the two confidence intervals have different interpretations. A nice feature of the posterior distribution is its interpretation. Unlike full Bayesian method where the posterior distribution is not connected to any tangible population, the posterior distribution of our method represents the distribution of RR among all drugs with $\widehat{RR}^* = 2.19$ and σ^*=0.345. In other words, the posterior distribution is calibrated with respect to $p(RR)$, subjects to potential bias due to estimating $\hat{p}(RR)$ (our simulation studies show that such a bias tends to be negligible for the data analyzed in this article) We will call the population of drugs with $\widehat{RR}^* = 2.19$ and σ^*=0.345 the *reference population* of the new drug.

Table 24 shows the posterior distribution evaluated at 1, 1.5, and 2. It can be seen that 97.3%, 43.6% and 9.7% of the drugs in the reference population have a true $RR > 1, 1.5$ and 2 (harmful effects). Similarly, 2.7%, 0% and 0% of the drugs in the reference population have a true $RR \leq 1, 0.67$ and 0.5 (protective effects). This is a particularly appealing interpretation of estimation uncertainty as it is natural and simple. The bias correction reduces the proportions to 91.2%, 26.7% and 4.9% for $RR > 1, 1.5$ and 2. Because we have limited number of drugs, the estimates of the posterior distribution demonstrate some level of variations as shown by the CIs in Table 24.

Table 24 Drug-outcome association inference from the posterior distribution ($\hat{p}\left(RR|Z^{*} = 2.272, \sigma^{*} = 0.345, \widehat{\Delta}^{*}\right)$)

Bias	$\Pr(RR \leq c)$			$\Pr(RR > c)$		
	$c = 1$	$c = 0.67$	$c = 0.5$	$c = 1$	$c = 1.5$	$c = 2$
	Estimate (95%)	Estimate (95%)	Estimate (95%)	Estimate (95%)	Estimate (95%)	Estimate (95%)
$\hat{\Delta}^{*} = 0$	0.027 [0.013,0.045]	0 (NA)	0 (NA)	0.973 [0.955,0.987]	0.436 [0.304,0.524]	0.097 [0.024,0.157]
$\hat{\Delta}^{*} = 0.31$	0.088 [0.056,0.14]	0 (NA)	0 (NA)	0.912 [0.86,0.944]	0.267 [0.156,0.346]	0.049 [0.007,0.096]

4.4 Simulation

We carried out two simulation studies. In simulation 1, we evaluated the level of accuracy in estimating $g(\delta)$ and its impact on $\hat{p}(RR|Z^*, \sigma^*)$, where we assumed there was no bias, i.e. $\Delta \equiv 0$, in data generation and in estimation. It is sufficient to study the case of $\Delta \equiv 0$ because we are primarily interested in how estimation of $g(\delta)$ affects posterior inference. In simulation 2, we assessed the combined impact of bias correction and the estimation of $g(\delta)$ on the inference based on $\hat{p}\left(RR|Z^*, \sigma^*, \hat{\Delta}^*\right)$. In this simulation, we generated bias for each drug according to some distribution during data generation and assumed $\hat{\Delta} = 0.31$ during estimation; the posterior distribution was conditioned on $\hat{\Delta}^* = 0$ or 0.31.

Simulation 1:

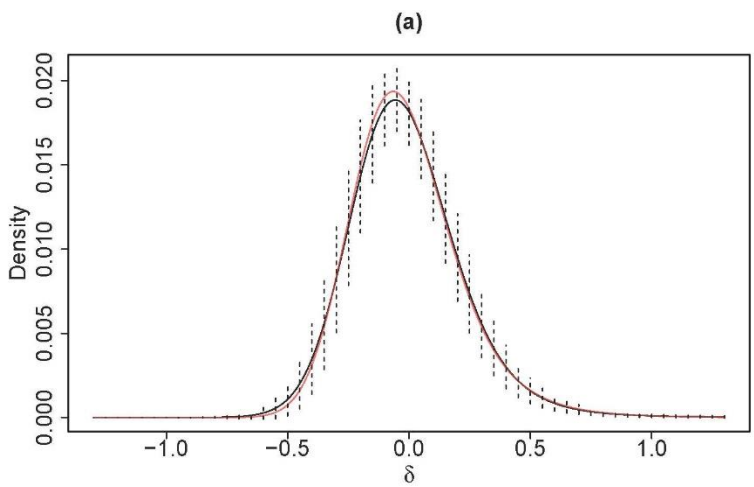
We simulated 500 Monte Carlo (MC) datasets. Each dataset included (Z_i, σ_i), $i = 1, 2, \ldots, 404$, generated by equation (3) with $\Delta \equiv 0$ (setting 1), where β_1, β_2 and g were set at the estimated values of section 0, and the empirical distribution of σ among the 404 drugs was assumed to be the true distribution of σ. Each MC data set was used to estimate β_1, β_2 and g. In addition, we generated $Z_j^*, j = 1, 2, 3$, for three exemplar new drugs whose relative risks are to be inferred from each MC data, with specific true risk ratio RR_j^*, standard error σ_j^*, and no bias $\Delta^* = 0$. Here RR_1^*, RR_2^* and RR_3^* were chosen to be close to the mean of $\hat{p}(RR)$ (with bias correction), the mean plus one standard deviation of $\hat{p}(RR)$ and the mean minus one standard deviation: $RR_1^* = 1, \sigma_1^* = 0.120$; $RR_2^* = 1.387$, $\sigma_2^* = 0.202$; $RR_3^* = 0.726$, $\sigma_3^* = 0.121$.

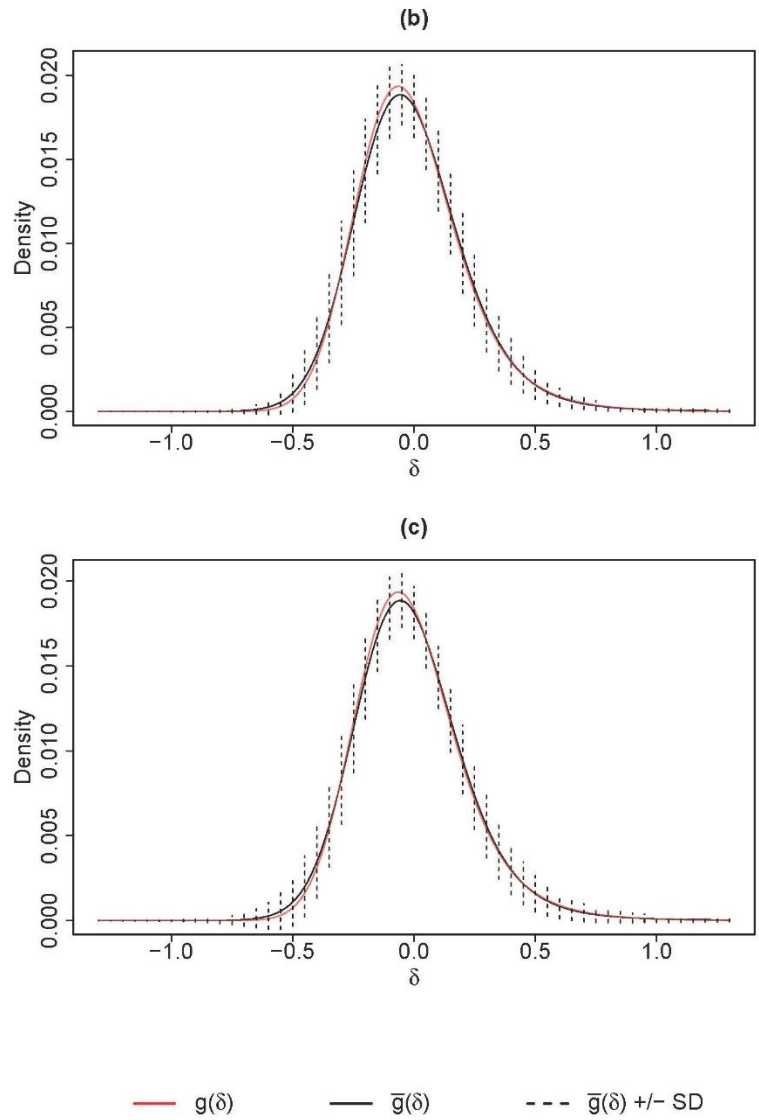
We applied the maximum likelihood estimation (MLE) to summary data (Z_i, σ_i) $i = 1, 2, \ldots, 404$ as described in Section 1.4 and obtained $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{g}$, which in turn were

used to estimate posterior $\hat{p}\left(\theta \middle| Z_j^*, \sigma_j^*\right)$. Furthermore, the true posterior $p\left(\theta \middle| Z_j^*, \sigma_j^*\right)$ was calculated based on true β_1, β_2 and g.

Figure 4(a) shows the mean of the estimates $\bar{g}(\delta_k) = \frac{\sum_{s=1}^{500} \hat{g}_s(\delta_k)}{500}$ in black and the true $g(\delta)$ in red at a set of grid points δ_k with a grid spacing of 0.01 $(k = 1, 2, \cdots, 261)$. The vertical dash bars represent $\bar{g}(\delta_k) \pm sd(\hat{g}(\delta_k))$. The plot shows that $\hat{g}$ is essentially unbiased at most locations except the neighborhood of the mode.

Figure 4 Density $\bar{g} \pm sd$ from 3 simulation settings compared with the true density g





4(a): $\Delta \equiv 0$ for both data generation and estimation.

4(b): $\Delta \sim N(0.31, \sigma^2)$, where $\sigma^2 = 0.13$. g is estimated by assuming a constant bias $\hat{\Delta} \equiv 0.31$.

4(c): $\Delta \sim N(0.31, 4\sigma^2)$, where $\sigma^2 = 0.13$. g is estimated by assuming a constant bias $\hat{\Delta} \equiv 0.31$.

To evaluate the performance of the estimated posterior distribution, we studied the coverage probability of the 95% credible region (CR), which is defined as the average

(over Monte Carlo simulations) probability mass between the 2.5 and 97.5 percentiles of $\hat{p}\left(RR\middle|Z_j^*, \sigma_j^*\right)$ under $p(RR|Z_i^*, \sigma_i^*)$. As shown in Table 25, the coverage probabilities for setting 1 are close to the nominal level.

Table 25 Coverage probability of the posterior distribution when $\widehat{\Delta}^* = \Delta^*$

Setting	Δ^*	$\widehat{\Delta}^*$	Coverage probability		
			$RR_1^* = 1$, $\sigma_1^* = 0.120$	$RR_2^* = 1.387$, $\sigma_2^* = 0.202$	$RR_3^* = 0.726$, $\sigma_3^* = 0.121$
1	0.00	0.00	95.2%	95.3%	94.6%
2	0.31	0.31	94.9%	94.9%	94.1%
3	0.31	0.31	93.9%	93.5%	91.6%

We summarized the bias and mean square error (MSE) of $\widetilde{RR}_j^*$, i.e. the mean of $\hat{p}\left(RR\middle|Z_j^*, \sigma_j^*\right)$. For comparison, we also computed the bias and the MSE of $\widehat{RR}_j^*$ ($\widehat{RR}_j^* = \exp\left(Z_j^* \sigma_j^*\right)$, which is the conventional frequentist estimate. As shown in Table 26 under setting 1, the bias of $\widetilde{RR}_j^*$ is acceptable and larger than that of $\widehat{RR}_j^*$ since $\widetilde{RR}_j^*$ is a shrinkage estimator[49]. The MSE of $\widetilde{RR}_j^*$ is smaller than that of $\widehat{RR}_j^*$ for $j = 1, 2$ because of the precision gain of $\widetilde{RR}_j^*$. The MSE of $\widetilde{RR}_3^*$ is greater than that of $\widehat{RR}_3^*$ since $\widehat{RR}_3^*$ already has a small standard error and the bias of the estimate dominates the MSE.

Table 26 Bias and MSE of $\widetilde{\mathrm{RR}}_{\mathrm{j}}^{*}$ from the posterior distribution when $\widehat{\Delta}^{*} = \Delta^{*}$

Setting	Δ^{*}	$\widehat{\Delta}^{*}$	Bayes posterior estimate $\widetilde{RR}_j^*$						$\widehat{RR}_j^*$					
			Bias			MSE			Bias			MSE		
			$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
1	0.00	0.00	0.025	-0.119	0.096	0.094	0.194	0.117	0.011	0.02	0.005	0.123	0.272	0.092
2	0.31	0.31	0.024	-0.111	0.089	0.096	0.195	0.113						
3	0.31	0.31	0.022	-0.086	0.071	0.102	0.200	0.103						

For reference, we have the true risk ratio and their standard error as following: $\mathrm{RR}_1^* = 1, \sigma_1^* = 0.120$; $\mathrm{RR}_2^* = 1.387, \sigma_2^* = 0.202$; $\mathrm{RR}_3^* = 0.726, \sigma_3^* = 0.121$.

Simulation 2:

In this simulation, data were generated the same way as in Simulation 1 except that a bias distribution was added. We considered two settings of bias distribution: $\Delta \sim N(0.31, \sigma^2)$ (setting 2) and $\Delta \sim N(0.31, 4\sigma^2)$ (setting 3), where $\sigma^2 = 0.13$. Here the value 0.13 was chosen because the variance of the bias should be equal to the variance of Z statistics for the negative controls (1.13) minus the variance in the absence of bias (1). For each of the three new drugs, we set the true bias at five values, $\Delta^* = 0.31 - 2\sigma, 0.31 - \sigma, 0.31, 0.31 + \sigma,$ and $0.31 + 2\sigma$.

$\hat{\beta}_1, \hat{\beta}_2$ and $\hat{g}$ were estimated as in Section 1.4 by assuming a constant bias $\hat{\Delta} \equiv 0.31$. The true posterior $p\left(\theta \middle| Z_j^*, \sigma_j^*, \Delta^*\right)$ was based on true β_1, β_2 and g, and the true bias Δ^*; the estimated posterior $\hat{p}\left(\theta \middle| Z_j^*, \sigma_j^*, \hat{\Delta}^*\right)$ was computed using $\hat{\beta}_1, \hat{\beta}_2$ and $\hat{g}$, and an assumed bias $\hat{\Delta}^*$, 0 or 0.31. Thus, in simulation 2, the assumed bias for each drug during estimation of model parameters and computation of posterior distribution is incorrect. When the assumed bias is 0.31, it at least matches the average bias. When the assumed bias is 0, it does not. Simulation 2 allows us to evaluate the impact of incorrect specification of the bias (together with the estimation of model parameters) on the posterior inference.

We assessed the estimation of $g(\delta)$ when an incorrect constant bias correction was used (Figure 4(b) and 4(c)). The plot shows that $\hat{g}$ is essentially unbiased at most locations except the neighborhood of the mode, the same feature as in Simulation 1. The 95% CR is the average (over Monte Carlo simulations) probability mass between the 2.5 and 97.5 percentiles of $\hat{p}\left(RR \middle| Z_j^*, \sigma_j^*, \hat{\Delta}^*\right)$ under $p(RR | Z_i^*, \sigma_i^*, \Delta^*)$. When $\hat{\Delta}^* = \Delta^*$, as

shown in Table 25, the coverage probabilities for setting 2 and 3 are close to the nominal level, the same as setting 1. When $\hat{\Delta}^* \neq \Delta^*$, as shown in Table 27, the coverage probabilities for setting 2 and 3 are greater than 90 % if the absolute difference between $\hat{\Delta}^*$ and Δ^* is small to moderate (less than 2σ).

Table 27 Coverage probability of the posterior distribution when $\widehat{\Delta}^* \neq \Delta^*$

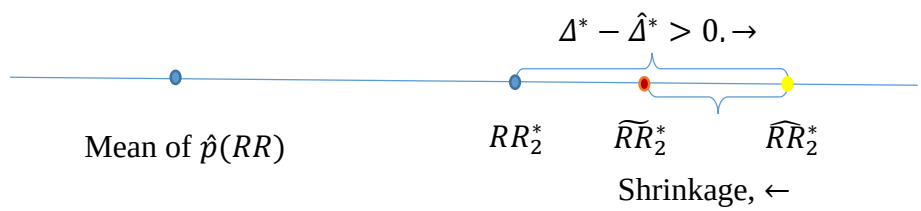
Setting	Δ^*	$\widehat{\Delta}^*$	Coverage probability		
			$RR_1^* = 1$, $\sigma_1^* = 0.120$	$RR_2^* = 1.387$, $\sigma_2^* = 0.202$	$RR_3^* = 0.726$, $\sigma_3^* = 0.121$
2	-0.408	0.310	90.5%	92.0%	89.0%
2	-0.408	0.000	94.3%	95.4%	93.1%
2	-0.049	0.310	93.8%	94.3%	92.5%
2	-0.049	0.000	95.0%	95.3%	94.1%
2	0.310	0.000	93.5%	93.4%	93.0%
2	0.669	0.310	93.6%	93.5%	93.5%
2	0.669	0.000	89.6%	89.6%	89.8%
2	1.028	0.310	90.0%	90.1%	90.7%
2	1.028	0.000	83.2%	83.4%	84.1%
3	-0.408	0.310	88.1%	90.6%	81.9%
3	-0.408	0.000	92.8%	94.1%	88.6%
3	-0.049	0.310	92.3%	93.1%	87.9%
3	-0.049	0.000	94.0%	93.8%	91.9%
3	0.310	0.000	92.8%	91.3%	92.7%
3	0.669	0.310	92.8%	91.4%	92.8%
3	0.669	0.000	88.8%	86.4%	90.9%
3	1.028	0.310	88.9%	86.8%	91.3%
3	1.028	0.000	81.9%	78.9%	86.3%

We summarized the bias and MSE of $\widetilde{RR}_j^*$, the mean of $\hat{p}\left(RR \middle| Z_j^*, \sigma_j^*, \hat{\Delta}^*\right)$. For comparison, we also computed the bias and MSE of $\widehat{RR}_j^*$, where $\widehat{RR}_j^* = \exp\left((Z_j^* - \hat{\Delta}^*)\sigma_j^*\right)$. When $\hat{\Delta}^* = \Delta^*$, as shown in Table 26, the bias and MSE of $\widetilde{RR}_j^*$ in setting 2 and

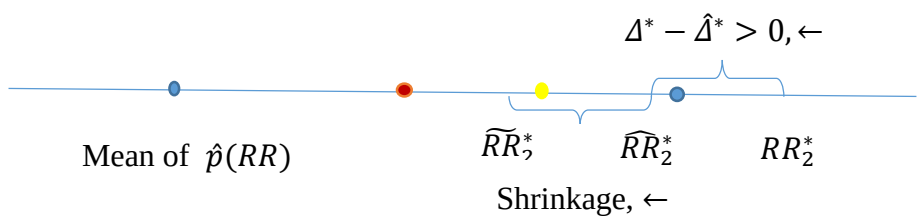
3 have the same features as in setting 1. When $\hat{\Delta}^* \neq \Delta^*$, both $\widetilde{RR}_j^*$ and $\widehat{RR}_j^*$ are affected by the incorrect bias correction. When $\Delta^* - \hat{\Delta}^* > 0$, the bias is under-corrected, pulling $\widetilde{RR}_j^*$ and $\widehat{RR}_j^*$ in the positive direction; When $\Delta^* - \hat{\Delta}^* < 0$, the bias is over-corrected, pulling $\widetilde{RR}_j^*$ and $\widehat{RR}_j^*$ in the negative direction.

We are mainly concerned about drugs with harmful effects $RR^* > 1$, e.g. in the case of testing the second drug. The effect of incorrect bias correction for $\widetilde{RR}_2^*$ and $\widehat{RR}_2^*$ and the shrinkage for $\widetilde{RR}_2^*$ are illustrated in Figure 5.

Figure 5 Combined effects of bias correction and shrinkage estimation on the bias of $\widetilde{\mathrm{RR}}_2^*$ compared with $\widehat{\mathrm{RR}}_2^*$



(a)



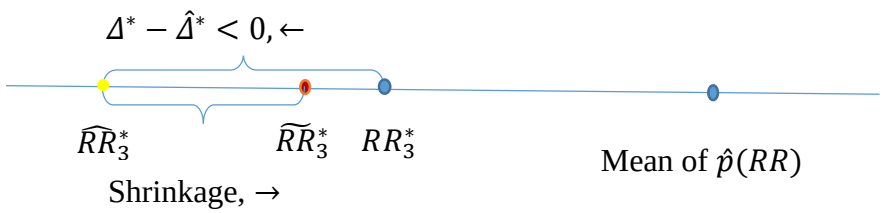
(b)

When $\Delta^* - \hat{\Delta}^* > 0$ (Figure 5(a)), the bias correction and shrinkage go in the opposite directions. Therefore the bias of $\widetilde{RR}_2^*$ is smaller than $\widehat{RR}_2^*$'s (e.g. Table 28, $\Delta^* = 0.31, \hat{\Delta}^* = 0$). When $\Delta^* - \hat{\Delta}^* < 0$ (Figure 5(b)), the bias correction and shrinkage go in

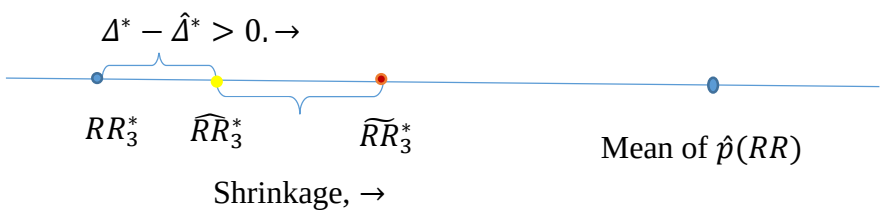
the same direction. Therefore the bias of $\widetilde{RR}_2^*$ is larger than $\widehat{RR}_2^*$'s (e.g. Table 28, $\Delta^* = -0.049, \hat{\Delta}^* = 0.31$).

For the third drug with protective effect ($RR^* < 1$), we illustrate the same phenomenon in Figure 6.

Figure 6 Combined effects of bias correction and shrinkage estimation on the bias of $\widetilde{\mathrm{RR}}_3^*$ compared with $\widehat{\mathrm{RR}}_3^*$



(a)



(b)

When $\Delta^* - \hat{\Delta}^* < 0$ (Figure 6(a)), the shrinkage and the bias correction go in the opposite directions so that the bias of $\widetilde{RR}_3^*$ is larger than that of $\widehat{RR}_3^*$ for most of the scenarios (e.g. Table 28, $\Delta^* = -0.408, \hat{\Delta}^* = 0$). When $\Delta^* - \hat{\Delta}^* > 0$ (Figure 6(b)), the shrinkage and the bias correction go in the same direction. Therefore the bias of $\widetilde{RR}_3^*$ is larger than that of $\widehat{RR}_3^*$ (e.g. Table 28, $\Delta^* = 0.31, \hat{\Delta}^* = 0$). When the shrinkage and the bias correction go in the opposite directions, the shrinkage could be so strong that it pulls

the $\widetilde{RR}_3^*$ to the other side of the true value, causing biases in $\widetilde{RR}_3^*$ and $\widehat{RR}_3^*$ with opposite signs (e.g. Table 28, $\Delta^* = -0.408, \hat{\Delta}^* = 0.31$).

Table 28 Bias and MSE of $\widetilde{RR}_j^*$ from the posterior distribution when $\widehat{\Delta}^* \neq \Delta^*$

Setting	Δ^*	$\widehat{\Delta}^*$	Bayes posterior estimate $\widetilde{RR}_j^*$						$\widehat{RR}_j^*$					
			Bias			MSE			Bias			MSE		
			$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
2	-0.408	0.310	-0.039	-0.218	0.043	0.094	0.256	0.077	-0.072	-0.171	-0.056	0.133	0.29	0.101
2	-0.408	0.000	-0.002	-0.143	0.072	0.089	0.201	0.098	-0.037	-0.092	-0.03	0.122	0.266	0.092
2	-0.049	0.310	-0.008	-0.167	0.065	0.090	0.222	0.094	-0.031	-0.079	-0.026	0.121	0.264	0.091
2	-0.049	0.000	0.030	-0.088	0.095	0.098	0.179	0.118	0.005	0.006	0.001	0.121	0.269	0.091
2	0.310	0.000	0.063	-0.028	0.120	0.116	0.174	0.140	0.049	0.111	0.034	0.136	0.309	0.101
2	0.669	0.310	0.058	-0.047	0.113	0.114	0.186	0.135	0.055	0.126	0.038	0.139	0.318	0.103
2	0.669	0.000	0.098	0.039	0.145	0.141	0.195	0.164	0.095	0.224	0.068	0.163	0.383	0.12
2	1.028	0.310	0.094	0.022	0.139	0.139	0.201	0.159	0.102	0.24	0.072	0.167	0.395	0.123
2	1.028	0.000	0.135	0.112	0.172	0.173	0.238	0.189	0.143	0.345	0.103	0.199	0.48	0.146
3	-0.408	0.310	-0.046	-0.208	0.022	0.102	0.257	0.071	-0.072	-0.171	-0.056	0.133	0.29	0.101
3	-0.408	0.000	-0.009	-0.133	0.049	0.095	0.208	0.088	-0.037	-0.092	-0.03	0.122	0.266	0.092
3	-0.049	0.310	-0.012	-0.150	0.045	0.096	0.223	0.085	-0.031	-0.079	-0.026	0.121	0.264	0.091
3	-0.049	0.000	0.025	-0.070	0.075	0.102	0.190	0.106	0.005	0.006	0.001	0.121	0.269	0.091
3	0.310	0.000	0.061	-0.003	0.101	0.120	0.192	0.128	0.049	0.111	0.034	0.136	0.309	0.101
3	0.669	0.310	0.059	-0.016	0.097	0.120	0.200	0.125	0.055	0.126	0.038	0.139	0.318	0.103
3	0.669	0.000	0.099	0.072	0.129	0.146	0.224	0.152	0.095	0.224	0.068	0.163	0.383	0.12
3	1.028	0.310	0.097	0.061	0.125	0.146	0.227	0.149	0.102	0.24	0.072	0.167	0.395	0.123
3	1.028	0.000	0.138	0.153	0.158	0.179	0.277	0.179	0.143	0.345	0.103	0.199	0.48	0.146

For reference, we have the true risk ratio and their standard error as following: $RR_1^* = 1, \sigma_1^* = 0.120$; $RR_2^* = 1.387, \sigma_2^* = 0.202$; $RR_3^* = 0.726, \sigma_3^* = 0.121$.

As shown in Table 28, the MSE of $\widetilde{RR}_j^*$ is smaller than that of $\widehat{RR}_j^*$ for $j = 1, 2$ due to the precision gain. But in most of the cases for $j = 3$, $\widetilde{RR}_3^*$ has a slightly larger MSE than $\widehat{RR}_3^*$. Since RR_3^* has a small standard error, the precision gain of $\widetilde{RR}_3^*$ cannot offset its bias and lead to the slightly larger MSE.

In summary, the simulation results show that the estimation of $g(\delta)$ is mostly accurate using our approach and that the true distribution of bias has little influence on it. All cases of $\widetilde{RR}_j^*$ $j = 1, 2$ have improved MSE through the precision gain. Although $\widetilde{RR}_3^*$ has a slightly inflated MSE in most of the cases, the drug has a protective effect and is of little safety concern.

4.5 Discussion

In this chapter, we focused on obtaining the distribution of the risk ratios so that the strength and direction of the association with a given outcome can be properly assessed for a population of drugs and an individual drug in the future. Our strategy includes a bias correction procedure leveraging drugs known to have no association with the outcome, followed by an empirical Bayes procedure based on the bias corrected risk ratio estimates.

Typical approaches for risk identification include the following designs and corresponding analytical methods: SCCS, new-user cohort design, case control design, self-controlled cohort design, temporal pattern discovery, and others (Ryan, Stang [50]). The main attractive feature of SCCS is that it is self-controlled, in which the effects of any fixed covariates cancel out. But there are also some limitations of which we need to be aware.

First, the validity of a risk ratio estimate by SCCS critically depends on the definitions of the outcome and the drug exposure. If a single event is mistakenly considered as multiple distinct events, e.g., subsequent treatment or management of an earlier AMI event are counted as multiple occurrences of AMI, such mistakes could lead to over- or under-estimation of the risk ratio. However, this can be overcome by considering the first event only for a patient. On the other hand, drug exposure could be under-captured since an EHR is by no means a closed system, where patients phase in and out due to moving or switching insurances, or have prescriptions filled outside the EHR. If they had been on a drug prior to their AMI but the drug exposure was not captured by the clinical database, we would under-estimate the log risk ratio (associated with false negatives).

Second, the success of using a patient's experience of AMI during the not-at-risk periods as his own reference is contingent on a key assumption that confounding covariates are time-invariant. Some covariates are certainly time-invariant, e.g., sex and race; however, age is not if age is indeed a confounder. It is possible to adjust for time-varying covariates in the estimation of risk ratio with by adding age as a covariate in the SCCS (Whitaker HJ, Farrington CP [45]).

Third, a patient's risk for an outcome may not be constant but diminish over time with continued use of a drug. The SCCS method may be more sensitive if the analysis is restricted to new users, and to immediate periods (duration is drug-dependent) rather than the entire observation period (defined as the entire period that a patient has been under observation in an EHR or the enrollment period for claims).

As most adverse drug events have low incidence, it usually requires a large number of subjects with exposure to the drug for the detection of these events with a sufficient statistical power. Thus, it is critical to develop efficient methods to detect signals as early and reliably as possible so that proper actions can be taken to avoid the occurrence of more events. Conventional analytical strategies that utilize only the data from the drug of interest for statistical evaluation are not efficient from this perspective. The empirical Bayes approach described in this article offers an objective, more efficient, and easily interpretable solution to this problem.

The estimation of the prior distribution is of interest for two main reasons. Firstly, we want to learn the ensemble properties such as $\Pr(RR > 2)$, which is an important metric to compute the number of drugs with strong associations with an adverse event. Secondly, we want to get the posterior distribution through empirical Bayes calculations, for instance, $\Pr(RR > 2|\widehat{RR}^* = 2.19, \sigma^* = 0.345)$. Compared with nonparametric methods from Laird [51], Fan [52], Hall and Meister [53] and Butucea and Comte [54], the Bayes deconvolution method we use here strikes a balance between convergence rate and robustness: the estimation has a root-*N* convergence rate and the class of the prior distributions as represented by splines is rich enough so it should be adequate for most practical applications. Accurate estimation of the prior distribution requires sufficient number of drugs. As shown in Table 23 and Table 24, there is some level of variations in the estimates of the prior/posterior quantities due to relatively limited number of drugs. As more and more data are accumulating in EHRs, more drugs may be included such that the variations can be substantially reduced to a level that is negligible. Thus, we believe

that our empirical Bayes approach holds strong potential for future post-marketing safety surveillance.

4.6 Conclusion

Bias correction can be used to improve inference accuracy for drug-outcome associations. The distribution of RR estimated through a large set of drug-outcome risk estimates provides an informative description of the state of drug-outcome associations and is readily useful for an empirical Bayes inference of the association for an individual drug-outcome pair.

APPENDIX A PROOF OF ASYMPTOTIC NORMALITY AND CONSISTENCY OF $\hat{\psi}_{MR}$

We first prove asymptotic normality and then prove consistency.

By definition

$$\hat{\psi}_{MR} \equiv \psi_{MR}(\bar{\mathbf{O}}; \hat{\alpha}_D(\hat{\gamma}, \hat{\omega}), \hat{\eta}_D(\hat{\gamma}, \hat{\omega}), \hat{\gamma}, \hat{\omega})$$

Let $\hat{\beta} \equiv (\hat{\alpha}_D(\hat{\gamma}, \hat{\omega})^T, \hat{\eta}_D(\hat{\gamma}, \hat{\omega})^T, \hat{\gamma}^T, \hat{\omega}^T)^T$ denote the vector of nuisance parameter estimates and $\beta^* \equiv (\alpha^{*,T}, \eta^{*,T}, \gamma^{*T}, \omega^{*,T})^T$ denote the limit of $\hat{\beta}$. If model (iii) for $\Pr(R = 1|Y, A, \mathbf{X})$ is correct, then γ^* is the true value γ_0. If model (iv) for $\Pr(Z = 1|R = 1, Y, A, \mathbf{X})$ or $E(Z|R = 1, Y, A, X)$ is correct, then ω^* is the true value ω_0. By the double robustness property of $\hat{\eta}_D(\hat{\gamma}, \hat{\omega})$, if either model (iii) and model (iv) is correct, and model (i) for $E[Y|A, \mathbf{X}, Z]$ is also correct, then η^* is the true value η_0. Similarly, by the double robustness property of $\hat{\alpha}_D(\hat{\gamma}, \hat{\omega})$, if either model (iii) and model (iv) is correct, and model (ii) for $\Pr(A = 1|\mathbf{X}, Z)$is also correct, then α^* is the true value α_0.

Let $S_\alpha(A, \mathbf{X}, Z)$, $S_\eta(Y, A, \mathbf{X}, Z)$, $S_\gamma(R, Y, A, \mathbf{X})$, and $S_\omega(Y, A, \mathbf{X}, Z)$ denote the score functions for the imposed working models $e(\mathbf{X}, Z; \alpha)$, $b(A, \mathbf{X}, Z; \eta)$, $\pi(Y, A, \mathbf{X}; \gamma)$, and $\varpi(Y, A, \mathbf{X}; \omega)$ respectively. Simple algebra leads to the following estimating equations (applying DR to the estimation of α and η):

$$0 = \mathrm{P}_n Q\left(\mathbf{O}; \hat{\beta}\right) \equiv \mathrm{P}_n \left\{ \begin{array}{c} \frac{R}{\pi(Y, A, \mathbf{X}; \hat{\gamma})} S_\alpha\left(A, \mathbf{X}, Z; \hat{\alpha}_D(\hat{\gamma}, \hat{\omega})\right) + \left(1 - \frac{R}{\pi(Y, A, \mathbf{X}; \hat{\gamma})}\right) \\ \times E\left[S_\alpha\left(A, \mathbf{X}, Z; \hat{\alpha}_D(\hat{\gamma}, \hat{\omega})\right) \middle| R = 1, Y, A, \mathbf{X}; \hat{\omega}\right] \\ \frac{R}{\pi(Y, A, \mathbf{X}; \hat{\gamma})} S_\eta\left(Y, A, \mathbf{X}, Z; \hat{\eta}_D(\hat{\gamma}, \hat{\omega})\right) + \left(1 - \frac{R}{\pi(Y, A, \mathbf{X}; \hat{\gamma})}\right) \\ \times E\left[S_\eta\left(Y, A, \mathbf{X}, Z; \hat{\eta}_D(\hat{\gamma}, \hat{\omega})\right) \middle| R = 1, Y, A, \mathbf{X}; \hat{\omega}\right] \\ S_\gamma(R, Y, A, \mathbf{X}; \hat{\gamma}) \\ S_\omega(Y, A, \mathbf{X}, Z; \hat{\omega}) \end{array} \right\}$$

Under suitable regularity conditions as stated in [16, 55, 56]

$$\sqrt{n}\left(\hat{\beta}-\beta^*\right)=\sqrt{n}\mathrm{P}_n\left\{\left[E\left(-\frac{\partial Q(\mathbf{O};\beta^*)}{\partial\beta^T}\right)\right]^{-1}Q(\mathbf{O};\beta^*)\right\}+o_p(1).$$

By Taylor expansion, we have

$$\sqrt{n}\left(\hat{\psi}_{MR}-\psi^*\right)=\sqrt{n}(\psi_{MR}(\overline{\mathbf{O}};\beta^*)-\psi^*)+\frac{\partial\psi_{MR}(\overline{\mathbf{O}};\beta^*)}{\partial\beta^T}\sqrt{n}\left(\hat{\beta}-\beta^*\right)+o_p(1),$$

where $\psi^*=E[\psi_{MR}(\overline{\mathbf{O}};\beta^*)]$. We could prove that $\psi^*=\psi_0$ if one of the four conditions from Section 0 holds. (Include more details)

Let $\psi_{MR}(\overline{\mathbf{O}};\beta)=\mathrm{P}_n\{\varphi_{MR}(\mathbf{O};\beta)\}$. Thus,

$$\begin{aligned}
&\sqrt{n}\left(\hat{\psi}_{MR}-\psi^*\right)\\
&=\sqrt{n}\mathrm{P}_n\{\varphi_{MR}(\mathbf{O};\beta^*)-\psi^*\}\\
&+\frac{\partial\psi_{MR}(\overline{\mathbf{O}};\beta^*)}{\partial\beta^T}\left(\sqrt{n}\mathrm{P}_n\left\{\left[E\left(-\frac{\partial Q(\mathbf{O};\beta^*)}{\partial\beta^T}\right)\right]^{-1}Q(\mathbf{O};\beta^*)\right\}+o_p(1)\right)\\
&+o_p(1)\\
&=\sqrt{n}\mathrm{P}_n\{\varphi_{MR}(\mathbf{O};\beta^*)-\psi^*\}\\
&+E\left[\frac{\partial\varphi_{MR}(\mathbf{O};\beta^*)}{\partial\beta^T}\right]\sqrt{n}\mathrm{P}_n\left\{\left[E\left(-\frac{\partial Q(\mathbf{O};\beta^*)}{\partial\beta^T}\right)\right]^{-1}Q(\mathbf{O};\beta^*)\right\}+o_p(1)\\
&=\sqrt{n}\mathrm{P}_n\left\{\varphi_{MR}(\mathbf{O};\beta^*)-\psi^*\right.\\
&\left.+E\left[\frac{\partial\varphi_{MR}(\mathbf{O};\beta^*)}{\partial\beta^T}\right]\left[E\left(-\frac{\partial Q(\mathbf{O};\beta^*)}{\partial\beta^T}\right)\right]^{-1}Q(\mathbf{O};\beta^*)\right\}+o_p(1)
\end{aligned}$$

By Central Limit Theorem, $\sqrt{n}\left(\hat{\psi}_{MR}-\psi^*\right)\to N(0,\Sigma)$, where

$$\Sigma=E\left[\varphi_{MR}(\mathbf{O};\beta^*)-\psi^*+E\left[\frac{\partial\varphi_{MR}(\mathbf{O};\beta^*)}{\partial\beta^T}\right]\left[E\left(-\frac{\partial Q(\mathbf{O};\beta^*)}{\partial\beta^T}\right)\right]^{-1}Q(\mathbf{O};\beta^*)\right]^2.$$

The covariance matrix Σ can be estimated by

$$\hat{\Sigma} = \mathrm{P}_n \left\{ \left(\varphi_{MR}(\mathbf{O};\hat{\beta}) - \hat{\psi}_{MR} + \mathrm{P}_n \left[\frac{\partial \varphi_{MR}(\mathbf{O};\hat{\beta})}{\partial \beta^T} \right] \left[\mathrm{P}_n \left(-\frac{\partial Q(\mathbf{O};\hat{\beta})}{\partial \beta^T} \right) \right]^{-1} Q(\mathbf{O};\hat{\beta}) \right)^2 \right\}$$

Next, we prove consistency, i.e., $\psi^* = \psi_0$ if one of conditions 1-4 listed in Section 0 holds. Note that $\psi^* = \mu_1^* - \mu_0^*$, where for $a \in \{0,1\}$ and $f_A(a, \mathbf{X}, Z; \alpha) = e(\mathbf{X}, Z; \alpha)^A \big(1 - e(\mathbf{X}, Z; \alpha)\big)^{1-A}$, we define

$$\begin{aligned}
&\mu_a^* \\
&= E\left[\frac{R}{\pi(Y, A, \mathbf{X}; \gamma^*)} \left(\frac{I(A=a)}{f_A(a, \mathbf{X}, Z; \alpha^*)} Y - \left(\frac{I(A=a)}{f_A(a, \mathbf{X}, Z; \alpha^*)} - 1 \right) b(A=1, \mathbf{X}, Z; \eta^*) \right) \right] \\
&- E\left[\left(\frac{R}{\pi(Y, A, \mathbf{X}; \gamma^*)} - 1 \right) \right. \\
&\times E\left[\left(\frac{I(A=a)}{f_A(a, \mathbf{X}, Z; \alpha^*)} - \left(\frac{I(A=a)}{f_A(a, \mathbf{X}, Z; \alpha^*)} - 1 \right) b(A=1, \mathbf{X}, Z; \eta^*) \right) \middle| R=1, Y, A, \mathbf{X}; \omega^* \right] \Bigg]
\end{aligned}$$

Without loss of generality, we prove that $\mu_1^* = E[Y_{a=1}]$ under conditional 1 in which model (i) for $E[Y|A, \mathbf{X}, Z]$ and model (iii) for $\Pr(R = 1|Y, A, \mathbf{X})$ are correct. Similar arguments can be made for other conditions, and then for μ_0^* and thus ψ^* under all 4 conditions.

As explained above, when models (i) and (iii) are correct, $\eta^* = \eta_0$ and $\gamma^* = \gamma_0$, then

$$\mu_{1,MR}(\beta^*)$$

$$= E\left[\frac{R}{\pi(Y,A,\mathbf{X};\gamma_0)}\left(\frac{I(A=1)}{e(\mathbf{X},Z;\alpha^*)}Y - \left(\frac{I(A=1)}{e(\mathbf{X},Z;\alpha^*)} - 1\right)b(A=1,\mathbf{X},Z;\eta_0)\right)\right]$$

$$- E\left[\left(\frac{R}{\pi(Y,A,\mathbf{X};\gamma_0)} - 1\right)\right.$$

$$\left.\times E\left[\left(\frac{I(A=1)}{e(\mathbf{X},Z;\alpha^*)} - \left(\frac{I(A=1)}{e(\mathbf{X},Z;\alpha^*)} - 1\right)b(A=1,\mathbf{X},Z;\eta_0)\right)\middle| R=1,Y,A,\mathbf{X};\omega^*\right]\right]$$

The second expectation is zero because $E\left[\left(\frac{R}{\pi(Y,A,\mathbf{X};\gamma_0)} - 1\right)|Y,A,X\right] = \frac{\pi(Y,A,\mathbf{X};\gamma_0)}{\pi(Y,A,\mathbf{X};\gamma_0)} - 1 = 0$.

Note that under MAR, $E[R|Y,A,\mathbf{X},Z] = E[R|Y,A,\mathbf{X}] = \pi(Y,A,\mathbf{X};\gamma_0)$, thus the first expectation equals

$$E\left[\frac{I(A=1)}{e(\mathbf{X},Z;\alpha^*)}Y - \left(\frac{I(A=1)}{e(\mathbf{X},Z;\alpha^*)} - 1\right)b(A=1,\mathbf{X},Z;\eta_0)\right]$$

$$= E\left[\frac{I(A=1)}{e(\mathbf{X},Z;\alpha^*)}\big(Y - b(A=1,\mathbf{X},Z;\eta_0)\big) + b(A=1,\mathbf{X},Z;\eta_0)\right]$$

$$= E[b(A=1,\mathbf{X},Z;\eta_0)] = E\big[E[Y|A=1,\mathbf{X},Z]\big] = E\big[E[Y_1|A=1,\mathbf{X},Z]\big]$$

$$= E\big[E[Y_1|\mathbf{X},Z]\big] = E[Y_1]$$

The second to last equality holds because of the assumption of no unmeasured confounding. The third to last equality holds because of the consistency assumption.

APPENDIX B AN EMPIRICAL VARIANCE ESTIMATE OF $\hat{\varphi}_a$ WITH BINARY MISSING COVARIATE Z

Assumptions:

A.1 Up to one from each of the two pairs of models (propensity and outcome regression, missing probability and $\varpi = f(Z = 1|Y, A, X)$ may be wrong so that $\hat{\varphi}_a$ is still consistent.

A.2 For models that are wrong, the parameter estimator converges to some value and is asymptotically normal.

We will use the same notation of $\theta = (\alpha, \eta, \varsigma, \gamma)$ to annotate the value the estimators converge to, with the understanding that values corresponding to the wrong models might not have clear meaning.

First note that for logistic regression:

$$\text{logit}[\Pr(Y = 1|X)] = \text{logit}(p_X) = X^T\beta$$

The score function is

$$S_\beta(Y) = X(Y - p_X)$$

Let E_* be the expectation with respect to $f(Z = 1|R = 1, Y, A, X)$. Based on Lingling's derivation:

$$\frac{\partial\hat{\varphi}_a}{\partial\varsigma} = -\mathbf{P}_n\left(\frac{R}{\pi^2}\frac{\partial\pi}{\partial\varsigma}(M_a - E_*(M_a))\right) = -\mathbf{P}_n\left(\frac{R}{\pi}(M_a - E_*(M_a))S_\varsigma(1)\right)$$

$$\frac{\partial\hat{\varphi}_a}{\partial\gamma} = -\mathbf{P}_n\left(\left(\frac{R}{\pi} - 1\right)E_*(M_a S_\gamma(1))\right)$$

$$\frac{\partial\hat{\varphi}_a}{\partial\alpha} = -\mathbf{P}_n\left(\frac{RI(A = a)}{\pi e_a}(Y - b_a)S_\alpha(a) + \left(1 - \frac{R}{\pi}\right)I(A = a)E_*\left(\frac{(Y - b_a)}{e_a}S_\alpha(a)\right)\right)$$

$$\frac{\partial\hat{\varphi}_a}{\partial\eta} = -\mathbf{P}_n\left(\frac{R}{\pi}\left(\frac{I(A = a)}{e_a} - 1\right)b_a S_\eta(a) + \left(1 - \frac{R}{\pi}\right)E_*\left[\left(\frac{I(A = a)}{e_a} - 1\right)b_a S_\eta(a)\right]\right)$$

Let $C_\bullet$ be the covariate vector associated with $\bullet$ ($\bullet = \alpha, \eta, \varsigma, \gamma$). The estimators $(\hat{\alpha}, \hat{\eta}, \hat{\varsigma}, \hat{\gamma})$ are determined by the following estimating equations:

$$\mathbf{P}_n Q = \mathbf{P}_n \begin{pmatrix} Q_1(\alpha, \varsigma, \gamma) \\ Q_2(\eta, \varsigma, \gamma) \\ Q_3(\varsigma) \\ Q_4(\gamma) \end{pmatrix} = \mathbf{P}_n \begin{pmatrix} \left(\frac{R}{\pi}(A - e_1)C_\alpha + \left(1 - \frac{R}{\pi}\right) E_*[(A - e_1)C_\alpha]\right) \\ \left(\frac{R}{\pi}(Y - b)C_\eta + \left(1 - \frac{R}{\pi}\right) E_*[(Y - b)C_\eta]\right) \\ (R - \pi)C_\varsigma \\ (Z - \varpi)C_\gamma \end{pmatrix} = 0$$

Under regularity conditions,

$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n} \begin{pmatrix} \hat{\alpha} - \alpha \\ \hat{\eta} - \eta \\ \hat{\varsigma} - \varsigma \\ \hat{\gamma} - \gamma \end{pmatrix} = \frac{1}{\sqrt{n}} \Sigma_n \left[E\left(-\frac{\partial Q}{\partial \theta^T}\right)\right]^{-1} Q + o_p(1)$ (Σ_n indicate summation over the n records)

$$\frac{\partial Q_1}{\partial \theta^T} = \begin{pmatrix} -\left(\frac{R}{\pi} e_1 C_\alpha S_\alpha^T(1) + \left(1 - \frac{R}{\pi}\right) E_*\left(e_1 C_\alpha S_\alpha^T(1)\right)\right), 0, \\ -\frac{R}{\pi}((A - e_1)C_\alpha - E_*((A - e_1)C_\alpha))S_\varsigma^T(1), \left(1 - \frac{R}{\pi}\right) E_*[(A - e_1)C_\alpha S_\gamma^T(1)] \end{pmatrix}$$

$$\frac{\partial Q_2}{\partial \theta^T} = \begin{pmatrix} 0, -\left(\frac{R}{\pi} b C_\eta S_\eta^T(1) + \left(1 - \frac{R}{\pi}\right) E_*(b C_\eta S_\eta^T(1))\right), \\ -\frac{R}{\pi}\left((Y - b)C_\eta - E_*((Y - b)C_\eta\right)S_\varsigma^T(1), \left(1 - \frac{R}{\pi}\right) E_*[(Y - b)C_\eta S_\gamma^T(1)] \end{pmatrix}$$

$$\frac{\partial Q_3}{\partial \theta^T} = (0,0, -\pi C_\varsigma S_\varsigma^T(1), 0)$$

$$\frac{\partial Q_4}{\partial \theta^T} = (0,0,0, -\varpi C_\gamma S_\gamma^T(1))$$

Based on Taylor expansion

$$\sqrt{n}\left(\hat{\varphi}_a(\hat{\theta}) - E(Y_a)\right)$$

$$= \sqrt{n}(\hat{\varphi}_a\ (\theta) - E(Y_a)) + \frac{\partial \hat{\varphi}_a}{\partial \alpha^T}\ \sqrt{n}(\hat{\alpha} - \alpha) + \frac{\partial \hat{\varphi}_a}{\partial \eta^T}\sqrt{n}(\hat{\eta} - \eta) + \frac{\partial \hat{\varphi}_a}{\partial \varsigma^T}\ \sqrt{n}(\hat{\varsigma} - \varsigma)$$

$$+ \frac{\partial \hat{\varphi}_a}{\partial \gamma^T}\ \sqrt{n}(\hat{\gamma} - \gamma) + o_p\ (1)$$

Let $\hat{\varphi}_a = P_n\ \hat{\phi}_a$

$$= \frac{1}{\sqrt{n}}\Sigma_n(\hat{\phi}_a(\theta) - E(Y_a)) + \left(\frac{\partial \hat{\varphi}_a}{\partial \alpha^T}, \frac{\partial \hat{\varphi}_a}{\partial \eta^T}, \frac{\partial \hat{\varphi}_a}{\partial \varsigma^T}, \frac{\partial \hat{\varphi}_a}{\partial \gamma^T}\right)\left(\frac{1}{\sqrt{n}}\Sigma_n\left[E\left(-\frac{\partial Q}{\partial \theta^T}\right)\right]^{-1} Q + o_p(1)\right)$$

$$= \frac{1}{\sqrt{n}}\Sigma_n(\hat{\phi}_a(\theta) - E(Y_a)) + E\left(\frac{\partial \hat{\phi}_a}{\partial \alpha^T}, \frac{\partial \hat{\phi}_a}{\partial \eta^T}, \frac{\partial \hat{\phi}_a}{\partial \varsigma^T}, \frac{\partial \hat{\phi}_a}{\partial \gamma^T}\right)\frac{1}{\sqrt{n}}\Sigma_n\left[E\left(-\frac{\partial Q}{\partial \theta^T}\right)\right]^{-1} Q + o_p(1)$$

$$= \frac{1}{\sqrt{n}}\Sigma_n\left[\hat{\phi}_a(\theta) - E(Y_a) + E\left(\frac{\partial \hat{\phi}_a}{\partial \alpha^T}, \frac{\partial \hat{\phi}_a}{\partial \eta^T}, \frac{\partial \hat{\phi}_a}{\partial \varsigma^T}, \frac{\partial \hat{\phi}_a}{\partial \gamma^T}\right)\left[E\left(-\frac{\partial Q}{\partial \theta^T}\right)\right]^{-1} Q\right] + o_p(1)$$

By CLT, the asymptotic variance of $\hat{\varphi}_a\ (\hat{\theta})$ can be estimated as

$$\widehat{Var}\left(\hat{\varphi}_a\ (\hat{\theta})\right) = \frac{1}{n} P_n \left[\hat{\phi}_a\ (\hat{\theta}) - \hat{\varphi}_a\ (\hat{\theta})\right.$$

$$\left. + \left[\left(\frac{\partial \hat{\varphi}_a}{\partial \alpha^T}, \frac{\partial \hat{\varphi}_a}{\partial \eta^T}, \frac{\partial \hat{\varphi}_a}{\partial \varsigma^T}, \frac{\partial \hat{\varphi}_a}{\partial \gamma^T}\right)\left(P_n\left(-\frac{\partial Q}{\partial \theta^T}\right)\right)^{-1}\right]\Big|_{\theta=\hat{\theta}} Q\right]^2$$

APPENDIX C AN EMPIRICAL VARIANCE ESTIMATE OF $\hat{\varphi}_a$ WITH CONTINUOUS MISSING COVARIATE Z

For the situation of continuous missing covariate, we substituted $\hat{z}$ for Z in the expectation part (for dealing with treatment effects). Thus, the estimator was as following:

$$\hat{\varphi}_a = \frac{R}{\pi} M_a - \left(\frac{R}{\pi} - 1\right) E_*(M_a)$$

where $M_a = \frac{A_a}{e_a} Y - \left(\frac{A_a}{e_a} - 1\right) b_a$, $E_*(M_a) = \frac{A_a}{\hat{e}_a} Y - \left(\frac{A_a}{\hat{e}_a} - 1\right) \hat{b}_a$. $A_a = I(A = a)$.

For $\widehat{e_a} = \Pr\left(A = 1 \middle| X, \hat{Z}\right) = expit\left(\alpha_0 + \alpha_1 X + \alpha_2 \hat{Z}\right) = expit[(\alpha_2\gamma_0 + \alpha_2\gamma_1 Y + \alpha_2\gamma_2 A + \alpha_2\gamma_3 X) + \alpha_0 + \alpha_1 X]$. We set $X' = \alpha_2(1\ Y\ A\ X)'$, $\theta = \alpha_0 + \alpha_1 X$ which was not related to γ. Then $\hat{e}_a = expit(X'\gamma + \theta) = \frac{\exp(X'\gamma+\theta)}{1+\exp(X'\gamma+\theta)}$.

$$\hat{l}_A(\gamma) = A \log \frac{\hat{e}_a}{1 - \hat{e}_a} + \log(1 - \hat{e}_a)$$

$$= A(X'\gamma + \theta) - \log[1 - \exp(X'\gamma + \theta)]$$

So we had $\hat{S}_{\gamma,A}(A) = \frac{\partial l_A}{\partial \gamma} = X'(A - \hat{e}_a)$.

$$\frac{\partial \widehat{e_a}}{\partial \gamma} = \frac{X' \exp(X'\gamma + \theta)}{(1 + \exp(X'\gamma + \theta))^2} = X'\hat{e}_a(1 - \hat{e}_a) = \hat{e}_a \hat{S}_{\gamma,A}(1)$$

Similarly, $\hat{b}_a = \Pr\left(Y = 1 \middle| A, X, \hat{Z}\right) = expit\left(\eta_0 + \eta_1 A + \eta_2 X + \eta_3 \hat{Z}\right)$

$$= expit[(\eta_3\gamma_0 + \eta_3\gamma_1 Y + \eta_3\gamma_2 A + \eta_3\gamma_3 X) + \eta_0 + \eta_1 A + \eta_2 X]$$

We set $X'' = \eta_3(1\ Y\ A\ X)'$, $\theta' = \eta_0 + \eta_1 A + \eta_2 X$. We had $\hat{S}_{\gamma,Y}(Y) = X''(Y - \hat{b}_a)$, $\frac{\partial \hat{b}_a}{\partial \gamma} = \hat{b}_a \hat{S}_{\gamma,Y}(1)$.

$$\frac{\partial \hat{\varphi}_a}{\partial \zeta} = -P_n\left[\frac{R}{\pi^2}\frac{\partial \pi}{\partial \zeta}\big(M_a - E_*(M_a)\big)\right] = -P_n\left[\frac{R}{\pi}\big(M_a - E_*(M_a)\big)S_\zeta(1)\right]$$

$$= -P_n\left[\frac{R}{\pi}\big(M_a - E_*(M_a)\big)(1-\pi)(1\ Y\ A\ X)'\right]$$

$$\frac{\partial \hat{\varphi}_a}{\partial \gamma} = -P_n\left[\left(\frac{R}{\pi}-1\right)\frac{\partial E_*(M_a)}{\partial \gamma}\right] = -P_n\left[\left(\frac{R}{\pi}-1\right)\frac{\partial}{\partial \gamma}\left(\frac{A_a}{\hat{e}_a}\big(Y-\hat{b}_a\big)+\hat{b}_a\right)\right]$$

$$= -P_n\left[\left(\frac{R}{\pi}-1\right)\left(-\frac{A_a}{\hat{e}_a}\big(Y-\hat{b}_a\big)\hat{S}_{\gamma,A}(1) - \frac{A_a}{\hat{e}_a}\hat{b}_a\hat{S}_{\gamma,Y}(1) + \hat{b}_a\hat{S}_{\gamma,Y}(1)\right)\right]$$

$$= -P_n\left[\left(\frac{R}{\pi}-1\right)\left(-\frac{A_a}{\hat{e}_a}\big(Y-\hat{b}_a\big)\hat{S}_{\gamma,A}(1) + \left(1-\frac{A_a}{\hat{e}_a}\right)\hat{b}_a\hat{S}_{\gamma,Y}(1)\right)\right]$$

$$= -P_n\left[\left(1-\frac{R}{\pi}\right)\left(\frac{A_a}{\hat{e}_a}\big(Y-\hat{b}_a\big)\hat{S}_{\gamma,A}(1) + \left(\frac{A_a}{\hat{e}_a}-1\right)\hat{b}_a\hat{S}_{\gamma,Y}(1)\right)\right]$$

$$= -P_n\left[\left(1-\frac{R}{\pi}\right)\left(\frac{A_a}{\hat{e}_a}\big(Y-\hat{b}_a\big)\alpha_2(1\ Y\ A\ X)'(a-\hat{e}_a)\right.\right.$$

$$\left.\left. + \left(\frac{A_a}{\hat{e}_a}-1\right)\hat{b}_a\eta_3(1\ Y\ A\ X)'\big(1-\hat{b}_a\big)\right)\right]$$

$$\frac{\partial \hat{\varphi}_a}{\partial \alpha} = P_n\left[\frac{\partial}{\partial \alpha}\left(\frac{R}{\pi}\frac{A_a}{e_a}(Y-b_a) + \left(1-\frac{R}{\pi}\right)\frac{A_a}{\hat{e}_a}\big(Y-\hat{b}_a\big)\right)\right]$$

$$= -P_n\left[\frac{R}{\pi}\frac{A_a}{e_a}(Y-b_a)S_\alpha(a) + \left(1-\frac{R}{\pi}\right)\frac{A_a}{\hat{e}_a}\big(Y-\hat{b}_a\big)\hat{S}_\alpha(a)\right]$$

$$= -P_n\left[\frac{R}{\pi}\frac{A_a}{e_a}(Y-b_a)(1\ X\ Z)'(a-e_a) + \left(1-\frac{R}{\pi}\right)\frac{A_a}{\hat{e}_a}\big(Y-\hat{b}_a\big)\big(1\ X\ \hat{Z}\big)'(a-\hat{e}_a)\right]$$

$$\frac{\partial \hat{\varphi}_a}{\partial \eta} = P_n\left[\frac{\partial}{\partial \eta}\left(\frac{R}{\pi}\left(1-\frac{A_a}{e_a}\right)b_a + \left(1-\frac{R}{\pi}\right)\left(1-\frac{A_a}{\hat{e}_a}\right)\hat{b}_a\right)\right]$$

$$= -P_n\left[\frac{R}{\pi}\left(\frac{A_a}{e_a}-1\right)b_aS_\eta^a(1) + \left(1\ \ -\frac{R}{\pi}\right)\left(\frac{A_a}{\hat{e}_a}-1\right)\hat{b}_a\hat{S}_\eta^a\ (1)\right]$$

$$= -P_n\left[\frac{R}{\pi}\left(\frac{A_a}{e_a}-1\right)b_a(1\ a\ X\ Z)'(1-b_a)+\left(1-\frac{R}{\pi}\right)\left(\frac{A_a}{\hat{e}_a}-1\right)\hat{b}_a\left(1\ a\ X\ \hat{Z}\right)'(1-\hat{b}_a)\right]$$

The estimators $\left(\hat{\alpha},\hat{\eta},\hat{\zeta},\hat{\gamma}\right)'$ were determined by the following equations:

$$P_nQ = P_n\begin{pmatrix} Q_1(\alpha,\zeta,\gamma) \\ Q_2(\eta,\zeta,\gamma) \\ Q_3(\zeta) \\ Q_4(\gamma) \end{pmatrix} = P_n\begin{pmatrix} \frac{R}{\pi}(A-e_1)C_\alpha + (1-\frac{R}{\pi})E_*[(A-e_1)C_\alpha] \\ \frac{R}{\pi}(Y-b)C_\eta + (1-\frac{R}{\pi})E_*[(Y-b)C_\eta] \\ (R-\pi)C_\zeta \\ R\left(Z-\hat{Z}\right)C_\gamma \end{pmatrix} = 0$$

Therefore, we have:

$$\frac{\partial Q_1}{\partial\theta^T} = \Bigg(-\left(\frac{R}{\pi}e_1C_\alpha S_\alpha^T(1)+(1-\frac{R}{\pi})E_*[e_1C_\alpha S_\alpha^T(1)]\right),\mathbf{0},-\frac{R}{\pi}((A-e_1)C_\alpha -$$

$$E_*[(A-e_1)C_\alpha])S_\zeta^T(1),(1-\frac{R}{\pi})\left(-\hat{e}_1\hat{C}_\alpha\hat{S}_{\gamma,A}(1)+(A-\hat{e}_1)\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ C_\gamma^T\end{pmatrix}\right)\Bigg)$$

$$\frac{\partial Q_2}{\partial\theta^T} = \Bigg(\mathbf{0},-\left(\frac{R}{\pi}bC_\eta S_\eta^T(1)+(1-\frac{R}{\pi})E_*[bC_\eta S_\eta^T(1)]\right),-\frac{R}{\pi}\Big((Y-b)C_\eta - E_*\big[(Y -$$

$$b)C_\eta\big]\Big)S_\zeta^T(1),(1-\frac{R}{\pi})\left(-\hat{b}\hat{C}_\eta\hat{S}_{\gamma,Y}(1)+\left(Y-\hat{b}\right)\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ C_\gamma^T\end{pmatrix}\right)\Bigg)$$

$$\frac{\partial Q_3}{\partial\theta^T} = \left(\mathbf{0},\mathbf{0},-\pi C_\zeta S_\zeta^T(1),\mathbf{0}\right)$$

$$\frac{\partial Q_4}{\partial\theta^T} = \left(\mathbf{0},\mathbf{0},\mathbf{0},-R\hat{Z}C_\gamma C_\gamma^T\right).$$

By CLT and Taylor expansion, the asymptotic variance of $\hat{\varphi}_a\left(\hat{\theta}\right)$ can be estimated as:

$$\widehat{Var}\left(\hat{\varphi}_a\left(\hat{\theta}\right)\right) = P_n\left\{\left(\hat{\phi}_a\left(\hat{\theta}\right)-\hat{\varphi}_a\left(\hat{\theta}\right)+P_n\left(\frac{\partial\hat{\varphi}_a}{\partial\alpha^T},\frac{\partial\hat{\varphi}_a}{\partial\eta^T},\frac{\partial\hat{\varphi}_a}{\partial\zeta^T},\frac{\partial\hat{\varphi}_a}{\partial\gamma^T}\right)\left[P_n\left(-\frac{\partial Q}{\partial\theta^T}\right)\right]^{-1}Q\right)^2\right\}$$

APPENDIX D OMOP ADJUDICATED LIST OF DRUGS POSITIVELY AND NEGATIVELY ASSOCIATED WITH AMI

Acute Myocardial Infarction				
Positive controls				
amlodipine	**ketorolac[48]**	almotriptan	factor VIIa	rizatriptan
darbepoetin alfa	**nabumetone[48, 49]**	amoxapine	fenoprofen[48]	salsalate
dipyridamole	**nifedipine**	bromocriptine	flurbiprofen[48]	tolmetin[48]
epoetin Alfa	**nortriptyline**	desipramine	frovatriptan	zolmitriptan
estradiol	**oxaprozin [48]**	diflunisal	imipramine	
estrogens, conjugated (USP)[50]	**piroxicam[48, 49, 51]**	eletriptan	ketoprofen[48, 49]	
etodolac[49, 52]	**sulindac[48, 51]**	enalaprilat	moexipril	
indomethacin[48, 49, 53, 54]	**sumatriptan**	estropipate	naratriptan	
Negative controls				
benzonatate	**ramelteon**	chlorothiazide	methenamine	stavudine
clindamycin	**salmeterol**	cosyntropin	methimazole	sulfasalazine
dicyclomine	**scopolamine**	darifenacin	miconazole	sulfisoxazole
fluticasone	**sitagliptin**	didanosine	nelfinavir	tetrahydrocannabinol
gatifloxacin	**sucralfate**	droperidol	nevirapine	thiabendazole
hyoscyamine	**temazepam**	endopeptidases	paromomycin	thiothixene
ketoconazole	**terbinafine**	entecavir	pemoline	tinidazole
lactulose	**urea**	ferrous gluconate	penicillamine	tipranavir
loratadine	acarbose	flavoxate	posaconazole	vitamin A
metaxalone	acetazolamide	flutamide	prilocaine	zafirlukast
methocarbamol	amylases	ketotifen	primidone	
penicillin V	bromfenac	lipase	propantheline	
prochlorperazine	chlorambucil	lithium citrate	simethicone	
oxybutynin	chlorazepate	mebendazole	sodiumphosphate, monobasic	

APPENDIX E BAYES DECONVOLUTION

We applied the MLE to obtain $\hat{g}(\delta)$ and $(\hat{\beta}_1, \hat{\beta}_2)$, where the sample space of δ is discretized at $(\delta_1, \dots, \delta_m)$ with grid precision 0.01[57]. Since σ and δ are independent, the estimated distribution of θ given σ is

$$\hat{p}(\theta|\sigma) = \hat{g}\left[\theta - \left(\hat{\beta}_1\sigma + \hat{\beta}_2\sigma^2\right)\right]. \tag{18}$$

Consequently,

$$\hat{p}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\hat{p}(\theta|\sigma_i) = \frac{1}{N}\sum_{i=1}^{N}\hat{g}\left[\theta - \left(\hat{\beta}_1\sigma_i + \hat{\beta}_2\sigma_i^2\right)\right] \tag{19}$$

Based on $\hat{g}(\delta)$, we employ Bayes rule to estimate the posterior distribution for a new drug with $\left(\widehat{RR}^*, \sigma^*\right)$. Let $\hat{\theta}^* = \log \widehat{RR}^*$. Then

$$\hat{p}\left(\theta\middle|\hat{\theta}^*, \sigma^*\right) \tag{20}$$

$$= \frac{\Pr\left(\hat{\theta}^*|\theta, \sigma^*\right) \cdot \hat{p}(\theta|\sigma^*)}{\sum_{j=1}^{m}\Pr\left(\hat{\theta}^*|\theta_j, \sigma^*\right) \cdot \hat{p}\left(\theta_j\middle|\sigma^*\right)}$$

$$= \frac{\Phi\left(\frac{\hat{\theta}^* - \theta}{\sigma^*}\right) \cdot \hat{g}\left(\theta - \left(\hat{\beta}_1\sigma^* + \hat{\beta}_2(\sigma^*)^2\right)\right)}{\sum_{j=1}^{m}\Phi\left(\frac{\hat{\theta} - \theta_j}{\sigma^*}\right) \cdot \hat{g}\left(\theta_j - \left(\hat{\beta}_1\sigma^* + \hat{\beta}_2(\sigma^*)^2\right)\right)}$$

Here $\theta_j = \delta_j + \hat{\beta}_1\sigma^* + \hat{\beta}_2(\sigma^*)^2, j = 1, \dots, m$, based on the δ's grid $(\delta_1, \dots, \delta_m)$. Based on $\hat{p}(\theta)$ and $\hat{p}\left(\theta|\hat{\theta}^*, \sigma^*\right)$, $\hat{p}(RR)$ and $\hat{p}\left(RR|\widehat{RR}^*, \sigma^*\right)$ can be derived through variable transformation.

BIBLIOGRAPHY

1. Birnbaum, H.G., et al., *Using healthcare claims data for outcomes research and pharmacoeconomic analyses.* PharmacoEconomics, 1999. **16** (1): p. 1-8.
2. Platt, R., et al., *Multicenter epidemiologic and health services research on therapeutics in the HMO Research Network Center for Education and Research on Therapeutics.* Pharmacoepidemiol Drug Saf, 2001. **10**(5): p. 373-377.
3. Sauer, B.C., et al., *Improving patient safety and pharmacovigilance: methods using observational data and cohort studies*. 2008, Agency for Healthcare Research and Quality.
4. Birnbaum, H.G., et al., *Using healthcare claims data for outcomes research and pharmacoeconomic analyses.* Pharmacoeconomics, 1999. **16**(1): p. 1-8.
5. Williamson, E., A. Forbes, and R. Wolfe, *Doubly robust estimators of causal exposure effects with missing data in the outcome, exposure or a confounder.* Statistics in medicine, 2012. **31**(30): p. 4382-4400.
6. Hernan, M.A., S. Hernandez-Diaz, and J.M. Robins, *A structural approach to selection bias.* Epidemiology, 2004. **15**(5): p. 615-25.
7. Little, R.J.A. and D.B. Rubin, *Statistical Analysis with Missing Data*. 1987, New York John Wiley & Sons.
8. Vansteelandt, S., A. Rotnitzky, and J. Robins, *Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse.* Biometrika, 2007. **94**(4): p. 841-860.
9. Kurth, T., et al., *Results of multivariate logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect.* American Journal of Epidemiology, 2005. **163** (3): p. 262-270.
10. Rosenbaum, P.R. and D.B. Rubin, *Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.* Journal of the American Statistical Association, 1984. **79**(387): p. 516-524.
11. Rosenbaum, P.R. and D.B. Rubin, *The Central Role of the Propensity Score in Observational Studies for Causal Effects.* Biometrika, 1983. **70**(1): p. 41-55.
12. Hernan, M.A., B. Brumback, and J.M. Robins, *Marginal structural models to estimate the causal effect of Zidovudine on the survival of HIV-positive men.* Epidemiology, 2000. **11** (5): p. 561-570.
13. Robins, J.M., M.A. Hernan, and B. Brumback, *Marginal structural models and causal inference in epidemiology.* Epidemiology, 2000. **11** (5): p. 550-560.
14. Bang, H. and J.M. Robins, *Doubly robust estimation in missing data and causal inference models.* Biometrics, 2005. **61** p. 962-972.
15. Robins, J.M., et al., *Higher order influence functions and minimax estimation of nonlinear functionals*, in *Probability and Statistics: Essays in Honor of David A. Freedman*, D. Nolan and T. Speed, Editors. 2008, Institute of Mathematical Statistics: Beachwood, Ohio. p. 335 -421.
16. Robins, J.M., A. Rotnitzky, and L.P. Zhao, *Analysis of Semiparametric Regression-Models for Repeated Outcomes in the Presence of Missing Data.* Journal of the American Statistical Association, 1995. **90**(429): p. 106-121.
17. Seymour, L., et al., *The Design of Phase II Clinical Trials Testing Cancer Therapeutics: Consensus Recommendations from the Clinical Trial Design Task*

Force of the National Cancer Institute Investigational Drug Steering Committee. Clinical cancer research : an official journal of the American Association for Cancer Research, 2010. **16**(6): p. 1764-1769.
18. Rubinstein, L., et al., *Randomized phase II designs.* Clin Cancer Res, 2009. **15**(6): p. 1883-90.
19. Pocock, S.J., *The combination of randomized and historical controls in clinical trials.* J Chronic Dis, 1976. **29**(3): p. 175-88.
20. Ibrahim, J.G. and M.-H. Chen, *Power prior distributions for regression models.* Statistical Science, 2000: p. 46-60.
21. Spiegelhalter, D.J., K.R. Abrams, and J.P. Myles, *Bayesian approaches to clinical trials and health-care evaluation*. Vol. 13. 2004: John Wiley & Sons.
22. Berry, S.M., et al., *Bayesian adaptive methods for clinical trials*. 2010: CRC press.
23. Neuenschwander, B., et al., *Summarizing historical information on controls in clinical trials.* Clin Trials, 2010. **7**(1): p. 5-18.
24. Hobbs, B.P., D.J. Sargent, and B.P. Carlin, *Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models.* Bayesian analysis (Online), 2012. **7**(3): p. 639.
25. Viele, K., et al., *Use of historical control data for assessing treatment effects in clinical trials.* Pharmaceutical statistics, 2014. **13**(1): p. 41-54.
26. O'Malley, A.J., S.-L.T. Normand, and R.E. Kuntz, *Sample size calculation for a historically controlled clinical trial with adjustment for covariates.* Journal of biopharmaceutical statistics, 2002. **12**(2): p. 227-247.
27. Pennello, G. and L. Thompson, *Experience with reviewing Bayesian medical device trials.* J Biopharm Stat, 2008. **18**(1): p. 81-115.
28. Bellingham, M.C., *A review of the neural mechanisms of action and clinical efficiency of riluzole in treating amyotrophic lateral sclerosis: what have we learned in the last decade?* CNS neuroscience & therapeutics, 2011. **17**(1): p. 4-31.
29. Bensimon, G., L. Lacomblez, and V.f. Meininger, *A controlled trial of riluzole in amyotrophic lateral sclerosis.* New England Journal of Medicine, 1994. **330**(9): p. 585-591.
30. Lacomblez, L., et al., *Dose-ranging study of riluzole in amyotrophic lateral sclerosis. Amyotrophic Lateral Sclerosis/Riluzole Study Group II.* Lancet, 1996. **347**(9013): p. 1425-1431.
31. Czaplinski, A., et al., *Predictability of disease progression in amyotrophic lateral sclerosis.* Muscle & nerve, 2006. **34**(6): p. 702-708.
32. Berry, J.D. and M.E. Cudkowicz, *New considerations in the design of clinical trials for amyotrophic lateral sclerosis.* Clinical investigation, 2011. **1**(10): p. 1375-1389.
33. Cudkowicz, M.E., et al., *Dexpramipexole versus placebo for patients with amyotrophic lateral sclerosis (EMPOWER): a randomised, double-blind, phase 3 trial.* The Lancet Neurology, 2013. **12**(11): p. 1059-1067.
34. Breslow, N.E. and J.A. Wellner, *Weighted Likelihood for Semiparametric Models and Two‐phase Stratified Samples, with Application to Cox Regression.* Scandinavian Journal of Statistics, 2007. **34**(1): p. 86-102.

35. Reid, N. and H. Crépeau, *Influence functions for proportional hazards regression.* Biometrika, 1985. **72**(1): p. 1-9.
36. Fleming, T.R. and D.P. Harrington, *Counting processes and survival analysis*. Vol. 169. 2011: John Wiley & Sons.
37. Investigators, W.G.f.t.W.s.H.I., *Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the Women's Health Initiative randomized controlled trial.* Jama, 2002. **288**(3): p. 321-333.
38. Bush, T.L., M. Whiteman, and J.A. Flaws, *Hormone replacement therapy and breast cancer: a qualitative review.* Obstetrics & Gynecology, 2001. **98**(3): p. 498-508.
39. Efron, B., *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. 2010: Cambridge University Press.
40. Schuemie, M.J., et al., *Interpreting observational studies: why empirical calibration is needed to correct p-values.* Stat Med, 2014. **33**(2): p. 209-18.
41. Diamond, G.A. and S. Kaul, *Prior convictions: Bayesian approaches to the analysis and interpretation of clinical megatrials.* Journal of the American College of Cardiology, 2004. **43**(11): p. 1929-1939.
42. Robbins, H., *Proc. Third Berkeley Symp. Math. Statist. Prob.* 1956.
43. Ryan, P., D. Griffin, and C. Reich, *OMOP common data model (CDM) Specifications.* 2009.
44. Ryan, P., *Establishing a drug era persistence window for active surveillance.* White papers, 2010.
45. Whitaker HJ, et al., *Tutorial in biostatistics: the self-controlled case series method.* Stat Med, 2006. **25**: p. 1768-1797.
46. Whitaker, H., *The self controlled case series method.* BMJ, 2008. **337**: p. a1069.
47. Ryan, P.B., et al., *Defining a reference set to support methodological research in drug safety.* Drug Saf, 2013. **36 Suppl 1**: p. S33-47.
48. Efron, B., *The Bayes deconvolution problem.* http://statweb.stanford.edu/~ckirby/brad/papers/2015BayesDeconvolutionProblem.pdf, 2015.
49. Efron, B., *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1. 2012: Cambridge University Press.
50. Ryan, P.B., et al., *A comparison of the empirical performance of methods for a risk identification system.* Drug safety, 2013. **36**(1): p. 143-158.
51. Laird, N., *Nonparametric maximum likelihood estimation of a mixing distribution.* J Am Stat Assoc, 1978. **73**(364): p. 805-811.
52. Fan, J., *On the optimal rates of convergence for nonparametric deconvolution problems.* The Annals of Statistics, 1991: p. 1257-1272.
53. Hall, P. and A. Meister, *A ridge-parameter approach to deconvolution.* The Annals of Statistics, 2007. **35**(4): p. 1535-1558.
54. Butucea, C. and F. Comte, *Adaptive estimation of linear functionals in the convolution model and applications.* Bernoulli, 2009. **15**(1): p. 69-98.
55. Robins, J.M. and A. Rotnitzky, *Semiparametric Efficiency in Multivariate Regression-Models with Missing Data.* Journal of the American Statistical Association, 1995. **90**(429): p. 122-129.

56. Robins, J.M., A. Rotnitzky, and L.P. Zhao, *Estimation of regression coefficients when some regressors are not always observed.* Journal of the American Statistical Association, 1994. **89** p. 846-866.
57. Efron, B., *The Bayes deconvolution problem.* 2015.

CURRICULUM VITAE

Jia Zhan

Education:

03/2017: Department of Biostatistics, Richard M. Fairbanks School of Public Health, Indiana University, Ph.D.

Major: Biostatistics Minor: Epidemiology

03/2006: Mathematics, Tongji University (China), M.S.

Major: Algebra-Lie algebra and representation Theory;

06/2003: Applied Mathematics, Tongji University (China), B.S.

Professional Experience:

01/2017-present: Principal Statistician at GlaxoSmithKline, Upper Providence, PA

07/2015-12/2016: Teach Assistant of Indiana University Public Health Biostatistics Department, Indianapolis, IN

01/2015-6/2015: Biostatistician Intern (Full time) at Biogen, Cambridge, MA

01/2013-12/2014: Research Assistant of Indiana University Public Health Biostatistics Department, Indianapolis, IN

03/2006-07/2011: Statistician of Information Center of Shanghai Disabled Persons' Federation, Shanghai, China

Awards:

2014 and 2015 Outstanding Advanced Graduate Student Awards By Indiana University Purdue University Indianapolis (IUPUI) School of Science

Publications:

1. Noll L. Campbell, Jia Zhan, Wanzhu Tu, Zach Weber, Roberta Ambeuhl, Caroline McKay, Newell McElwee. Self-reported medication adherence barriers among ambulatory older adults with mild cognitive impairment. Pharmacotherapy. 2016.
2. YE JC, Jia Zhan. The Opinion on Textbook Reform of Linear Algebra. College Mathematics. 2006;2:004.
3. YE JC, Ben Sheng Y, Jia Zhan. A New Multiplicity Formula for the Weyl Modules of Type B and C. Paper presented at: Journal of Mathematical Research & Exposition 2010.