Developing and Evaluating Innovative Approaches for Estimating Causal Effects of Low-dose Aspirin on Pregnancy Outcomes

by

Yongqi Zhong

BMed, Central South University, China, 2015

MPH, Washington University in St. Louis, United States, 2017

Submitted to the Graduate Faculty of

the Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Yongqi Zhong

It was defended on

July 1, 2021

and approved by

Maria M. Brooks, PhD

Professor of Epidemiology and Biostatistics, University of Pittsburgh

Sofia Triantafyllou, PhD

Assistant Professor of Biomedical Informatics, University of Pittsburgh

Edward H. Kennedy, PhD

Associate Professor of Statistics and Data Science, Carnegie Mellon University

Dissertation Co-Director: Lisa M. Bodnar, PhD, MPH, RD

Professor of Epidemiology, University of Pittsburgh

Dissertation Co-Director: Ashley I. Naimi, PhD

Associate Professor of Epidemiology, Emory University

Developing and Evaluating Innovative Approaches for Estimating Causal Effects of Low-dose Aspirin on Pregnancy Outcomes

Yongqi Zhong, PhD

University of Pittsburgh, 2021

First trimester pregnancy loss occurs in one third of all pregnancies, and recurrent pregnancy loss is also prevalent in up to 30% of women with a prior history. Using intention-to-treat, the Effects of Aspirin in Gestation and Reproduction (EAGeR) trial found that low-dose aspirin (LDA) led to 4.3 (95% CI -1.2 to 9.6) per 100 women at high risk of pregnancy loss. However, the estimated effect, which is based on the assignment to a treatment arm, rather than adherence to a particular treatment protocol, limits the understanding of potential benefits of LDA on pregnancy.

Existing methods for adherence adjustment to estimate per-protocol effects in randomized trials are subject to the limitations of observational studies, including model mis-specification due to incorrect confounder selection, or from strong parametric assumptions.

The objective of this dissertation is to evaluate and develop innovative approaches for estimating the adherence-adjusted effects of LDA on pregnancy. First, to mitigate the impact of incorrect confounder selection, we evaluated the performance of causal discovery methods in a simulation study using the data resampled from EAGeR. We found that, the evaluated causal discovery method yielded low accuracy in selecting sufficient confounder adjustment sets in the M- or Butterfly-structured causal diagrams. Second, to avoid strong parametric assumptions, we developed an R package implementing the augmented inverse probability weighting (AIPW), a doubly robust estimator supporting

stacking machine learning. Our simulation study suggests that, our **AIPW** package has excellent performance compared to existing R packages implementing doubly robust estimators. Finally, we used the **AIPW** package with stacking machine learning to estimate per-protocol effects of LDA in a time-fixed setting from the EAGeR trial. Our results show that LDA led to 8.0 (95% CI 2.5 to 13.6) more pregnancies per 100 women who adhered to the randomized treatment assignment for at least 5/7 days per week over at least 80% person-week of follow-up, consistent with the previous analysis using parametric g-formula in a time-varying setting. In conclusion, this dissertation does not only provide additional evidence of the benefits of LDA on pregnancy, but also the state-of-the-art approaches for effect estimations in epidemiologic studies.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

1.1 Specific Aims

First trimester pregnancy loss occurs in one third of all pregnancies, and recurrent pregnancy loss is also prevalent in up to 30% of women with a prior history. Preconception low-dose aspirin has shown promises in preventing adverse pregnancy outcomes, likely due to its pro-circulatory and anti-inflammatory effects that influence critical conception process (e.g. implantation). The Effects of Aspirin in Gestation and Reproduction (EAGeR) trial estimated an increase of 4.3% in the pregnancy rate (95% CI -1.2% to 9.6%) among women assigned to preconception low-dose aspirin versus placebo. However, the estimated effect, which is based on the assignment of treatment rather than adherence, limits the understanding of potential benefits of low-dose aspirin to improving pregnancy outcomes. Non-adherence, as well as post-randomization confounding can lead to large differences in estimated treatment effects, because randomization only accounts for eliminating or reducing the imbalance of baseline characteristics among treatment arms. To disentangle such differences, not only does one require solid understanding of the causal structure of low-dose aspirin and pregnancy outcomes, but one must also correctly specify all statistical models used to model these causal relations (e.g. what variables to be adjusted and which statistical methods to be used). Unfortunately, uncertainty about the precise causal structure of low-dose aspirin on pregnancy outcome and its correctly specified statistical models leaves a major gap of understanding in the true biological effects of low-dose aspirin on pregnancy outcomes.

In the long term, our work will develop a modern framework and tools with advanced

methodology for identifying effective treatments to improve pregnancy outcomes. Our overall objective is to develop and evaluate causal inference tools to estimate the adherence adjusted effect of low-dose aspirin on pregnancy. We hypothesize that low-dose aspirin will increase the pregnancy rate after adjusting for non-adherence and post-randomization confounding. We will test our hypothesis with the EAGeR study, a multicenter randomized trial of the effect of daily low-dose aspirin on multiple pregnancy outcomes (e.g. live birth, pregnancy loss). From 2007 to 2012, this trial recruited 1,228 women (18-40 years) with a lifetime history of one or two pregnancy losses, who are attempting to become pregnant. Participants were followed for up to six menstrual cycles (for whom did not conceive) or throughout pregnancy.

Aim 1. Evaluate the performance of Bayesian networks in selecting covariate adjustment sets for estimating the effects of low-dose aspirin on pregnancy outcomes.

Widely applied in computer science and bioinformatics, Bayesian networks are a quantitative representation of causal diagrams used to discover and model complex causal structures. Using simulation studies, we will first evaluate the performance of different covariate adjustment in estimating the effects of low-dose aspirin on pregnancy outcomes (e.g. adjusting the covariates that are statistical related but not the cause of the exposure and outcome) with the knowledge of true data generating mechanisms. Then, we will evaluate the performance of Bayesian networks in identifying the true causal structure and selecting the best optimal adjustment sets for effect estimations.

Aim 2. Develop an R package for augmented inverse probability weighting (AIPW) to estimate the effects of low-dose aspirin on pregnancy outcomes.

Methods in Aim 1 guides the use of covariate adjustment sets for modeling the mechanism of taking low-dose aspirin and the mechanism relating to pregnancy outcomes.

However, it is not typically possible to correctly specify all statistical models given investigators' uncertainty about the true data generating mechanisms. The AIPW, a doubly robust estimator, is able to provide a valid estimation of causal effects as long as either the outcome or the exposure model is correctly specified. We will develop a new R package for AIPW which is robust to an extent of statistical model mis-specifications, and compare the performance of our package to existing packages that also implement doubly robust estimators.

Aim 3. Determine the adherence adjusted effects of low-dose aspirin on pregnancy outcomes with Bayesian networks and AIPW.

We will estimate the effect of adherence to preconception low-dose aspirin use on the pregnancy outcomes by adjusting for post-randomization confounding. First, we will use the methods evaluated in Aim 1 to select a covariate adjustment set reflecting the causal structure among low-dose aspirin, pregnancy outcomes, and factors relating to non-adherence. Second, we will use the AIPW package developed in Aim 2 to estimate the adherence adjusted effects and compare to other commonly used methods (e.g. g-computation).

In completing these aims, this study will provide better understanding of the performance of Bayesian networks and AIPW in estimating the causal relationships among low-dose aspirin, pregnancy outcomes and factors leading to non-adherence. Not only will this potentially inform the complex causal relationships between low-dose aspirin and pregnancy loss, but we will also provide a more comprehensive framework and tools guiding the use of graphical models and other advanced analytical techniques in reproductive epidemiological studies to prevent adverse pregnancy outcomes.

1.2 Significance

Burden of adverse pregnancy outcomes increases over the past decades in the United States.

As an adverse event affecting women's physical and mental health, pregnancy loss is a common complication among conceived women—one third of women in their early pregnancy had pregnancy loss (or miscarriage).[1, 2] According to the CDC, risk of pregnancy loss increased 1% per year among US women from 1990 to 2011. Pregnancy loss is also considered as a risk factor for future fertility.[3] In addition to pregnancy loss and fertility, the prevalence of preterm birth in the US shows an increasing trend from 2014 to 2016—about 1 out of 10 infants were born before 37 weeks of gestation.[4] Thus, more research on adverse pregnancy outcomes prevention is needed.[5]

Preconception low-dose aspirin is promising to prevent adverse pregnancy outcomes.

Low-dose aspirin, a commonly used, cheap, over-the-counter medication, is promising to prevent pregnancy loss for its pro-circulatory and anti-inflammatory effects. The effect of postconceptional use of low-dose aspirin on early pregnancy loss has been studied over the past decades, whose evidence is mixed.[6, 7, 8, 9, 10, 11] However, postconceptional use of low-dose aspirin is still being prescribed for pregnancy loss prevention without solid evidence. Preconception low-dose aspirin, on the other hand, shows more promises than postconceptional low-dose aspirin, since preconception low-dose aspirin could directly influence the critical windows of pregnancy (e.g. implantation) so as to improve pregnancy rate and to prevent adverse pregnancy outcomes.

Discrepancies in different estimated treatment effects limit the potential benefits of low-dose-aspirin on pregnancy outcomes.

The Effects of Aspirin in Gestation and Reproduction (EAGeR) is a landmark randomized controlled trial (RCT), which was among the first to examine the effect of preconception LDA on pregnancy outcomes in women with one or two prior pregnancy losses. The study investigators found an increased probability of 4.3% in pregnancy (95% CI -1.2% to 9.6%) among women assigned to preconception low-dose aspirin versus placebo.[12, 13, 14, 15] This estimated intention-to-treat effect was unable to yield sufficient statistical evidence to change the clinical guideline. However, non-adherence was a documented problem in EAGeR, suggesting that the per-protocol effect may suffice to provide solid statistical evidence. Our work will address the differences in estimated treatment effects of low-dose aspirin on pregnancy outcomes.

Non-adherence may lead to large differences in the estimated treatment effects in RCTs.

Intention-to-treat analysis uses treatment assignment as the exposure for treatment effect estimation, regardless of whether a subject is adherent to the assigned treatment or not; whereas the per-protocol analysis is based on adherence status in addition to the treatment assignment. Ideally, the estimated intention-to-treat effect is identical to per-protocol effect because all participants were assumed to adhere to the assigned treatment throughout the trial.[16] However, because of non-adherence, the estimation of intention-to-treat effect might not reflect the clinical effectiveness of the treatment.[17] The magnitude and direction of changes in intention-to-treat effect depend on the type of trial and the pattern of non-adherence[18, 19] For example, intention-to-treat effect could underestimate the actual treatment effect in the safety and non-inferiority trial, leading to observe the equivalent or similar effect between the treatment and the placebo. This underestimation of intention-to-treat effect in non-inferiority trial would yield conservative evidence that the proposed treatment is as "effective" as the traditional one.[19] There

are three commonly used methods for correcting non-adherence in estimating treatment effect:[20] instrumental variable methods,[21] principal stratification[22] and marginal structural modelling.[23] All of these methods require solid understanding of variables which strongly predict the outcome and which tightly link to non-adherence—that is, post-randomization confounding.

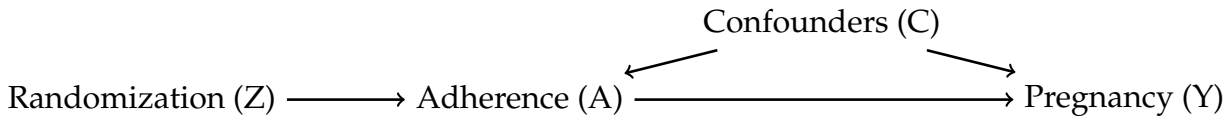
Post-randomization confounding is critical for non-adherence adjustment in estimating per-protocol effects in RCTs.

One major advantage of RCTs over observational studies is randomization, which, in expectation, ensures exchangeability between treatment and control groups at baseline.[24] As a result, the intention-to-treat effect is identified and unconfounded. However, the per-protocol effect might be confounded as in observational studies,[25] because randomization only accounts for eliminating the imbalance at the time of randomization and does not guarantee potential violations of exchangeability for the per-protocol effect estimation.[26] Using EAGeR as an example, although assigning to aspirin or placebo group did not depend on baseline prognostic covariates (e.g. employment status), whether adhering to the treatment assignment may be affected by those covariates (e.g., older age may increase the chance of taking aspirin but decrease the chance of being pregnant). Analyzing RCTs with similar analytical approaches and considerations as observational studies had been recently advocated and successfully applied in some clinical studies.[17, 25, 27]

Causal diagrams provide novel insights to address non-adherence and post-randomization confounding in RCTs.

As a routinely used tool by epidemiologists, causal diagrams allow investigators to evaluate the identifiability of treatment effect with their background knowledge and to

Figure 1: A causal diagram for the EAGeR trial



select a set of covariates used for adjustment.[24, 28, 29, 30] For example, Figure 1 depicts the simplified causal relations in an RCTs like EAGeR, where Z denotes treatment assignment to aspirin by randomization, A is the adherence to assigned treatment, Y is the outcome and C is the post-randomization confounding; as such, the intention-to-treat effect can be estimated by $P(Y \mid Z)$. However, the per-protocol effect cannot be directly estimated by $P(Y \mid Z, A = 1)$ because it is confounded by C, similar to the targeted effects in observational studies. Therefore, we need to estimate $P(Y \mid Z, C, A = 1)$ to unbiasedly estimate the effects of adherence with aspirin on pregnancy. In empirical studies, C in Figure 1 could potentially include multiple post-randomization confounders, which may be associated with each other. Take EAGeR as an example, side effects and symptoms such as bleeding are post-randomization confounders, which are also associated with each other, holding an intercorrelated causal structure in addition to the adherence-outcome relations. Without causal diagrams, the potential problems introduced by post-randomization confounders cannot be easily evaluated. Therefore, the choice of which method to use for analysis (e.g., traditional regression versus more advanced g methods) and which variables to adjust for cannot easily be determined.[28]

Graphical models provide potential solutions to address the complexity of modeling causal relations,[24, 28, 31] such as identifying the covariate adjustment sets.[32, 33, 34] Our work will evaluate and develop methods relating to these modern causal inference

methodologies (Aim 1 & 2).

Bayesian networks are able to model the complex casual relations relating to preconception low-dose aspirin and pregnancy outcomes.

While numerous variable selection algorithms exist,[35] such as step-wise regression, the majority of them do not incorporate causal thinking, which could potentially select a covariate that introduce bias (e.g., collider).[32, 36, 37] Developed and widely applied in machine learning and bioinformatics, Bayesian networks are often used to construct a complex system (e.g. medical diagnostic system).[38, 39, 40] As a quantitative representation of the causal diagram, Bayesian networks are a multivariate probabilistic graphical model that can also be used to model and identify causal relations.[41] Unfortunately, in population health science, few has used the broader framework offered by Bayesian networks—causal discovery (the process whereby data are used to uncover potential causal structures with minimal background knowledge).[42, 43] Further, since these casual discovery algorithms are developed and in bioinformatic data (e.g., normalized gene expressions),[39, 42, 43] their performance are unknown in selecting covariate adjustment sets in realistic epidemiologic setting, which contains various types of data (e.g. continuous and categorial) with small to moderate effect sizes. Evaluating the performance of Bayesian networks would inform our modelling the complex causal relationship among low-dose aspirin, pregnancy outcomes and factors relating to non-adherence in the EAGeR trial. Our Aim 1 will set a foundation for using causal discovery algorithms to select an optimal covariate adjustment set for treatment effect estimation in RCT and observational studies.

Doubly robust estimators are robust to the bias introduced by statistical model misspecifications.

After identifying a casual diagram, investigators can use methods evaluated in our Aim 1 to estimate the treatment effect of interest. Take Figure 1 as an example to estimate the per-protocol effect [$P(Y \mid Z, A = 1)$], IPW requires the estimates from the exposure model [i.e., $P(Z \mid C, A = 1)$], and g-computation requires the estimates from the outcome model [i.e., $P(Y \mid Z, C, A = 1)$].[24, 29, 31] Specifying the exposure or the outcome model requires knowledge of which variables to be included in the model and what type of the statistical model to be used.[44, 45] Notably, parametric regressions are often used to model the exposure and outcome, which require more statistical assumptions, such as normality.[46] Violating those assumptions results in mis-specifying statistical models, and further introduces bias in treatment effect estimations. Alternatively, machine learning methods are becoming population for estimating the exposure and outcome models because they require few parametric assumptions. However, recent theoretical and empirical studies suggest that estimating treatment effects with machine learning methods should be accompanied by doubly robust estimators.[46, 47, 48, 49, 50] Doubly robust estimators combine the estimates from both exposure and outcome models.[51, 52] As a result, as long as one of the exposure or the outcome model is correctly specified, the doubly robust estimators are able to yield consistent treatment effect estimations, providing a "second" chance for model specifications (i.e., the doubly robust property).[47, 51, 52, 53]

As a doubly robust estimator, augmented inverse probability weighting (AIPW)[51] has a number of implementations in different programming languages,[54, 55, 56] which are often estimated by parametric models (e.g. linear/logistic regressions). Unfortunately, only a handful of the programs implementing AIPW enable use of machine learning methods and few of them can be tailored for RCTs. Our Aim 2 will fill this gap by developing a state-of-the-science R package. Not only will this package will equip us to

estimate the adherence adjusted effects of low-dose aspirin on pregnancy outcome, but also will allow epidemiologists to investigate the causal effects of other exposures to improve adverse pregnancy outcomes in both observational studies and RCTs.

Models adjusting for adherence reveal the biological effect of low-dose aspirin on pregnancy outcomes.

Because of non-adherence, the biological effect of preconception low-dose aspirin is underestimated by the intention-to-treat effect reported in the EAGeR.[17, 25, 27] Previous publications of EAGeR found that preconception low-dose aspirin increases of pregnancy rate;[12] and further, conception also led to a 10% decrease in adherence among EAGeR participants. Based on the drop of adherence rate after conception, we hypothesize that low-dose aspirin could increase pregnancy rate after adjusting for non-adherence and post-randomization confounders. With the methods and tools developed in Aim 1 & 2, our Aim 3 will determine the time-fixed effects of preconception low-dose aspirin on pregnancy in the EAGeR trial.[15]

1.3 Overall Impact

Successful completion of this work will 1) yield better understanding of the graphical models (e.g. Bayesian networks) in selecting covariate adjustment set for estimating the causal effects relating to adverse pregnancy outcomes; 2) provide a modern statistical program (the AIPW package) for effect estimations in both observational studies and randomized trials in reproductive epidemiology; and 3) reveal the complex causal relationships among low-dose aspirin, pregnancy outcomes and factor relating to non-adherence by using graphical models and doubly robust programs. These will further serve as a

more informed framework for using modern causal inference methods to improve populational reproductive health.

1.4 Innovation

Existing research on aspirin and pregnancy outcomes has been dominated by studies that report the intention-to-treat effect only with parametric models (e.g., logistic regression),[6, 7, 8, 9, 10, 57, 12, 13, 16] which limits the understanding of the biological effects of aspirin on pregnancy outcomes. Although intention-to-treat effect has been a gold standard for clinical trials,[16] it underestimates the biological effect because of non-adherence.[18, 19] In addition, the statistical assumptions of parametric models may lead to model mis-specifications, introducing bias for the treatment effect estimations.[46, 49] Our proposed work is highly innovative:

1. Rarely applied in population health setting, Bayesian networks facilitate identifying the causal relationships among low-dose aspirin, pregnancy outcomes and factors relating to non-adherence, providing in-depth insights on the complex causal structures for per-protocol estimations.[41]
2. Tailored for randomized trials, the AIPW package will be flexible and robust to model mis-specifications, assuring that the per-protocol effect are appropriately estimated with adjustment of non-adherence and post-randomization confounders.[52, 53, 58]
3. Machine learning methods are data-adaptive so as to avoid the model mis-specifications introduced by the assumptions from parametric models.[46, 48, 49] Machine learning methods have been suggested to apply with doubly robust estimators. The AIPW package will be implemented with cross-fitting, statistical technique to address the

issues in estimating treatment effects with machine learning methods (e.g. high bias, high mean square error).[46, 47, 48, 49, 50]

2.0 Challenges in automating confounder selection with causal discovery methods

2.1 Introduction

Directed acyclic graphs (DAGs) have now become a cornerstone of many epidemiologic analyses. Epidemiologists often use DAGs to conceptualize the causal relationships relevant to the exposure-outcome relation of interest.[59, 60] When interest lies in estimating the effect of an exposure on an outcome, DAGs, along with a set of rules for reading DAGs (such as d-separation), can be used to identify covariates that should be included or excluded from the adjustment set to minimize bias in the exposure-outcome effect estimate.[61, 59, 32, 33, 34] General recommendations are to adjust for pre-treatment covariates,[33, 34, 62] except when there is a strong belief that such adjustment would introduce or amplify bias (e.g., M-bias, or Z-bias).[33, 34, 36, 37, 63, 64]

However, correct identification of a confounder set that eliminates bias requires correct specification of the underlying DAG. Given that true causal relations underlying the data are rarely (if ever) known, it is not typically possible to identify a single set of covariates that suffice to yield unbiased causal effect estimates. In most settings, domain-specific knowledge may be compatible with a wide array of causal diagrams, leading to completely different confounder adjustment sets. Furthermore, even if the true causal relations were known, uncertainty may arise when choosing between a covariate set that is minimally sufficient [61, 59, 33, 34] and a set that includes covariates that might improve statistical efficiency. [62, 65]

Though not commonly employed in epidemiology, DAGs can be used in other ways. Instead of relying on domain-specific knowledge to generate a DAG, and using d-separation

to identify an adjustment set for *causal inference*,[61, 59, 32, 33, 34] one may attempt to use data to generate a DAG in a process of *causal discovery*.[66, 67, 61, 39, 68, 43] With minimal background knowledge, causal discovery proceeds by testing the graphical properties of input data and trying to recover causal relations of the variables in the dataset. [61, 39, 43] In principle, use of causal discovery algorithms can yield insights on the causal relations underlying a given dataset. These insights can then be used for several purposes, including a preliminary selection of covariates to adjust for confounding in the process of causal inference.[69, 70] Indeed, such procedures have already been employed in other disciplines where domain-specific knowledge is limited.[71]

In this paper, we illustrate the process of causal discovery, and evaluate the performance of causal discovery algorithms in correctly selecting covariates adjustment sets. Additionally, since average treatment effects (ATEs) are often of interest to epidemiologists, we examine the bias in estimating ATEs using plasmode simulations constructed to resemble realistic epidemiologic scenarios. We evaluate the performance of causal discovery algorithms in conjunction with a range of estimators for the ATE.

2.2 Methods

2.2.1 Data Generating Mechanisms

We simulated data to explore the performance of different causal discovery algorithms using data from the Effects of Aspirin in Gestation and Reproduction (EAGeR) trial.[15] The data generating mechanism in Figure 2 was used to create scenarios with M- and Butterfly-biases, where adjusting for the collider alone (i.e., C_1 in Figure 2) would intro-

duce bias.[61, 36, 37, 64] M-bias and butterfly-bias were selected because they entail particularly difficult challenges for causal discovery algorithms, and have been implicated in realistic epidemiologic settings.[72, 36]

The EAGeR study is a multicenter randomized trial of the effect of daily low-dose aspirin on pregnancy outcomes in 1,228 women aged 18-40 years with one or two prior pregnancy loss, who were attempting to become pregnant. We resampled $1,228$ observations from the EAGeR baseline data, consisting of the randomized treatment assignment indicator (Z), number of prior pregnancy losses (C_2), and high sensitivity C reactive protein (hsCRP, denoted as C_3). We also resampled a set of discrete, continuous, and ordinal covariates denoted B, which are meant to represent variables in the dataset that are indirectly associated with the causal system under study through an unmeasured variable U. These variables B were included as correlated nuisance variables that are not of any relevance to the effect of interest, but that may influence the performance of causal discovery algorithms. Characteristics of these variables are shown in Table 1. Details about the EAGeR trial are available elsewhere.[12, 14, 15]

With these resampled variables, we use the following three logistic regression models (Equation 1-3) to simulate a dichotomous exposure A (e.g., adherence to aspirin in EAGeR), outcome Y (e.g., live birth) and collider/confounder C_1, with marginal probabilities of $0.25, 0.5, 0.5$, respectively.[73] The parameters of these logistic regression models were chosen to yield a DAG as displayed in Figure 2.

Table 1: Baseline covariates from the EAGeR trial used for simulations

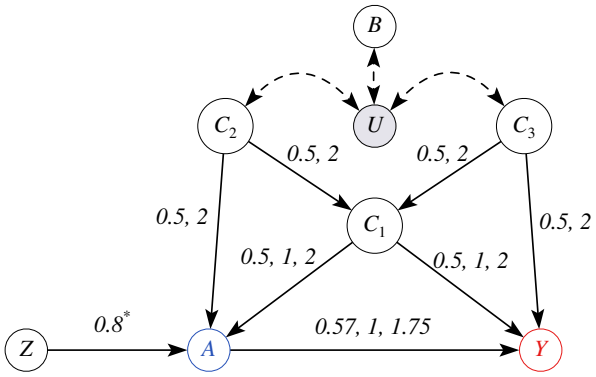
Variables	Values	n (%) / mean (SD)
$\mathbf{C}_2$: Num. previous pregnancy losses (%)	1	825 (67.2)
	2	403 (32.8)
$\mathbf{C}_3$: hsCRP level (%)	Low: < 2 mg/L	785 (63.9)
	Medium: $[2, 10)$ mg/L	377 (30.7)
	High: ≥ 10 mg/L	66 (5.4)
$\mathbf{Z}$: Randomized treatment assignment (%)	Low-dose Aspirin=1	615 (50.1)
	Placebo=0	613 (49.9)
$\mathbf{B}$:		
Eligibility stratum (%)	New	679 (55.3)
Alcohol use in the past year (%)	Never	820 (66.8)
	Sometimes	382 (31.1)
	Often	26 (2.1)
Smoking in the past year (%)	Yes	152 (12.4)
Employed (%)	Yes	920 (74.9)
Age (mean (SD))		28.74 (4.80)
Months of conception attempts prior to randomization (mean (SD))		4.03 (3.45)
BMI (mean (SD))		26.32 (6.57)
Mean arterial pressure (mean (SD))		85.54 (9.55)

$$\text{logit}[P(C_1 = 1 \mid C_2, C_3)] = \beta_{C_{10}} + \beta_{C_1C_2}C_2 + \beta_{C_1C_3}C_3 \tag{1}$$

$$\text{logit}[P(A = 1 \mid Z, C_1, C_2)] = I(Z = 1)(\beta_{A_0} + \beta_{AZ}Z + \beta_{AC1}C_1 + \beta_{AC_2}C_2) \tag{2}$$

$$\text{logit}[P(Y = 1 \mid A, C_1, C_3)] = \beta_{Y_0} + \beta_{YA}A + \beta_{YC1}C_1 + \beta_{YC_3}C_3 \tag{3}$$

Figure 2: A causal diagram for butterfly (M) bias with an instrumental variable



- Numbers on the edges represent the ORs between variables
- $P(A = 1 \mid Z = 1, C_1, C_2) \geq 0$ and $P(A = 1 \mid Z = 0, C_1, C_2) = 0$

where $\beta = log(OddsRatio)$, $\beta_{AZ} = log(0.8)$.

This data generating mechanism yields a dataset with one instrument (Z), one exposure (A), one outcome (Y), three covariates for the M-/Butterfly- structure ($C_1 \dots C_3$) and eight correlated nuisance variables (B). To explore the impact of different strength and directions of the associations, we set the odds ratios (ORs) of $\{0.5, 2\}$ for the M structure of the DAG (i.e., $\{C_2 \to A, C_2 \to C_1, C_3 \to C_1, C_3 \to Y\}$), $\{0.5, 1, 2\}$ for the exposure-collider-outcome relations (i.e., $\{C_1 \to A, C_1 \to Y\}$), and $\{0.57, 1, 1.75\}$ for the exposure-outcome relation (i.e., $\{A \to Y\}$). These associations are shown in Figure 2. The combinations of associations among the simulated variables resulted in a total of $4^2 \cdot 3^2 \cdot 3 = 432$ data generating mechanisms that could be classified into four types of structures: 192 butterfly-structure DAGs; 48 M-structure DAGs; 96 butterfly-structure DAGs without an arrow between C_1 and Y (or left-triangle structure), and 96 butterfly-structure DAGs without an

arrow between C_1 and A (or right-triangle structure). For each scenario, we analyzed a total of 2,000 Monte Carlo (MC) datasets, each with a sample size of 1,228 observations, yielding a total of 864,000 datasets explored.

2.2.2 Statistical Analysis Plan

2.2.2.1 Average treatment effect

We used the data generated above to estimate the average treatment effect of A on Y. The average treatment effect (ATE) is defined as the average outcome if all observations were exposed versus unexposed. This effect can be quantified on the risk difference scale as:

$$ATE_{RD} = E(Y^1 = 1) - E(Y^0 = 1)$$

where Y^1 and Y^0 are the potential outcomes that would be observed if the exposure (A) was set to 1 and 0, respectively.

Under exchangeability, consistency, no interference, and positivity, the average of potential outcomes can be quantified as $E(Y^a) = E[E(Y|A = a, C)]$, which is the estimated mean of the observed outcomes among those with $A = a$, averaged over the observed distribution of C. This estimand is often quantified with inverse probability weighting (IPW), g-computation (sometimes referred to as the parametric g formula), augmented inverse probability weighting (AIPW), or targeted maximum likelihood estimation (TMLE).[74, 75, 76, 77] Details about these estimators are provided in the Appendix A.1.

2.2.2.2 Manual confounder adjustment set selection

Under the causal diagram depicted in Figure 2, several different adjustment sets can yield an unconfounded estimator of the ATE.[61, 59, 37] These are confounder adjustment sets containing C_1 with C_2 and/or C_3 for the butterfly-structure, any sets other than C_1 only for the M-structure, sets containing C_3 or $\{C_1, C_2\}$ for left-triangle-structure, and sets containing C_2 or $\{C_1, C_3\}$ for right-triangle-structure DAGs (Table 2).

Table 2: Confounder adjustment sets that block the backdoor paths from A to Y

DAG type	Num. DAGs	Admissible adjustment sets[1]
Butterfly ⋈	192	$\boldsymbol{\{C_1, C_2\}, \{C_1, C_3\}, \{C_1, C_2, C_3\}}$
M ⋀	48	Any[2] except $\{C_1\}$
Left-triangle ⋈	96	$\{C_3\}$, $\{C_2, C_3\}$, $\boldsymbol{\{C_1, C_2\}, \{C_1, C_3\}, \{C_1, C_2, C_3\}}$
Right-triangle ⋈	96	$\{C_2\}$, $\{C_2, C_3\}$, $\boldsymbol{\{C_1, C_2\}, \{C_1, C_3\}, \{C_1, C_2, C_3\}}$

[1] Adjustment sets with nuisance variables B are not shown in this table.
[2] Empty adjustment set $\{\emptyset\}$ / unadjusted estimate is admissible.
Bold sets are common admissible adjustment for all 432 DAGs.

Based on our knowledge of the true data generating mechanism, we manually choose four possible confounder adjustment scenarios to evaluate the absolute bias and mean squared error (MSE) of each estimator under different settings. These four scenarios are: adjusting for direct causes of the exposure and the outcome, which are $\{C_1, C_2, C_3\}$ in butterfly, left-triangle, right-triangle DAGs, and $\{C_2, C_3\}$ for M-structured DAGs (i.e., all causes with collider scenario in Table 3); adjusting for nuisance variables B in addition to the all cause scenario (i.e., all covariates scenario in Table 3); adjusting for the collider C_1 only, which would introduce bias in the M-structure data generating mechanism or in the structures that C_1 does not fully block the backdoor path (e.g., in the butterfly-structure);[32, 36, 37, 64] adjusting for an empty set of covariates, which would provide

unadjusted estimates.

Table 3: Covariate adjustment scenarios for estimating average treatment effects in the M-/butterfly-structure causal diagram

Scenarios	Adjustment Set
Manual Covariate Selection	
All Causes (with Collider)[✓]	$\{C_1,$[1] $C_2, C_3\}$
All Covariates[✓]	$\{C_1, C_2, C_3, B\}$
Collider Only	$\{C_1\}$
Empty Set / Unadjusted	$\{\emptyset\}$
Automated Covariate Selection by Causal Discovery[2]	
All Causes[3]	Direct cause(s) of A and/or Y

[1] C_1 was not adjusted for in M-structured DAGs.
[2] Default and tuned MMHC algorithms (conditional independence test = mutual information with $\alpha = 0.05, 0.1$) with minimal prior knowledge for restrictions (i.e., forced edges: $Z \rightarrow A \rightarrow Y$ and forbidden edges: covariates pointing to Z or Z pointing to covariates and Y). Instrumental variable Z was not adjusted.
[3] Accuracy of correct covariate adjustment set is evaluated with causes of A and/ Y
[✓] Adjustment set that is guaranteed to block the backdoor path from A to Y for all 432 DAGs

2.2.2.3 Automated confounder adjustment set selection with causal discovery

Defining a DAG with domain-specific knowledge is equivalent to assuming that the causal structure represents the mechanism that generated the data. Causal discovery is a data-driven approach that tries to recover the causal structure by identifying the DAG(s) that "fit" the data (e.g., detecting dependencies among variables). [66, 67, 61, 39, 68, 43] Here, we use the Max-Min Hill Climbing (MMHC) algorithm in the simulated data above in an attempt to identify the relevant confounder adjustment sets automatically.[78] The MMHC has been shown to perform well in settings similar to epidemiologic data,[70, 79]

and is relatively straightforward to use via bnlearn R package.[80] A brief introduction to the MMHC causal discovery algorithm is provided in Appendix A.2.

We explore the performance of the MMHC algorithm with two parameter settings (default and tuned). The default setting involves conditional independence tests with a significance level of $\alpha = 0.05$ to construct the skeleton of the DAG, and a score function defined via the Bayesian Information Criterion (BIC). The tuned setting is identical to the default setting, with the exception of setting α to 0.1 for a denser graph.[60] For the purpose of confounder selection, several preprocessing steps before fitting the MMHC algorithm should be used. These include first categorizing continuous data into quintiles, applying restrictions to the DAG based on reasonable background knowledge (e.g., forcing edges $Z \to A \to Y$, and forbidding other edges from pointing into the instrumental variable Z and from pointing into the candidate confounders from the exposure and the outcome).

In this automated confounder selection process, we use the direct causes of exposure and outcome as the confounder adjustment set (all causes scenario in Table 3). These causes are identified in the DAGs generated by the MMHC algorithms. The original data were used for ATE estimation, instead of the discretized data that were used to fit the MMHC algorithms.

2.2.3 Performance Evaluation

To evaluate the accuracy of confounder selection, we define the correct confounder adjustment set as the direct causes of exposure and outcome (all cause scenario in Table 3) that blocks any backdoor paths. With this definition, accuracy is computed as the number of correctly selected confounder adjustment sets divided by the total number of

scenarios. For more details about the accuracy, we further calculate the probability of the MMHC algorithms in selecting C_1, C_2, C_3, and confounder adjustment sets using their combinations (i.e., $\binom{3}{0} + \binom{3}{1} + \binom{3}{2} + \binom{3}{3} = 8$ sets), respectively. Similarly, this probability is defined as the number of selected covariate/confounder adjustment set divided by the number of DAGs * 2000 MC.

We also evaluate the impact of different confounder adjustment scenarios in Table 3 on the performance of ATE estimation using g-computation, IPW, AIPW and TMLE. The correctly specified parametric regression model for the data generating mechanisms was used to generate 1 million observations to obtain the estimates of the true parameter value (ATE_{true}). For each data generating mechanism of the total 432 DAGs, absolute bias and mean squared error (MSE) are computed, defined as $Abs.Bias = E(|\widehat{ATE} - ATE_{true}|)$ and $MSE = E[(\widehat{ATE} - ATE_{true})^2]$, respectively. To simplify the analyses, if the MMHC algorithms yielded an empty confounder adjustment set, unadjusted estimates (i.e., $\hat{P}(Y = 1|A)$) were used instead of g-computation, IPW, AIPW and TMLE. Among all 432 DAGs, median with 25% and 75% quartiles (Q1, Q3) of the absolute bias and MSE are reported, stratified by confounder selection scenarios. To compare the median (Q1,Q3) of absolute bias across scenarios, we compute a relative absolute bias using the all cause scenario from manual selection as the reference, defined as the median (Q1,Q3) of each scenario divided by the median (Q1,Q3) of the manually selected all cause setting. Distributions of the absolute bias and MSE for the 432 DAGs are visualized in the Appendix A.5, with the strata of manual and automated selections, types of DAG and the four estimators.

Simulations, ATE estimation, and causal discovery were conducted in R (3.6.0) on a Linux-based computing cluster supported by the University of Pittsburgh Center for Research Computing, and analyses of the performance were conducted in R (3.6.2) on a

local Apple Computer (Catalina 10.15.7).[81]

2.3 Results

2.3.1 Accuracy of causal discovery algorithms in confounder selection

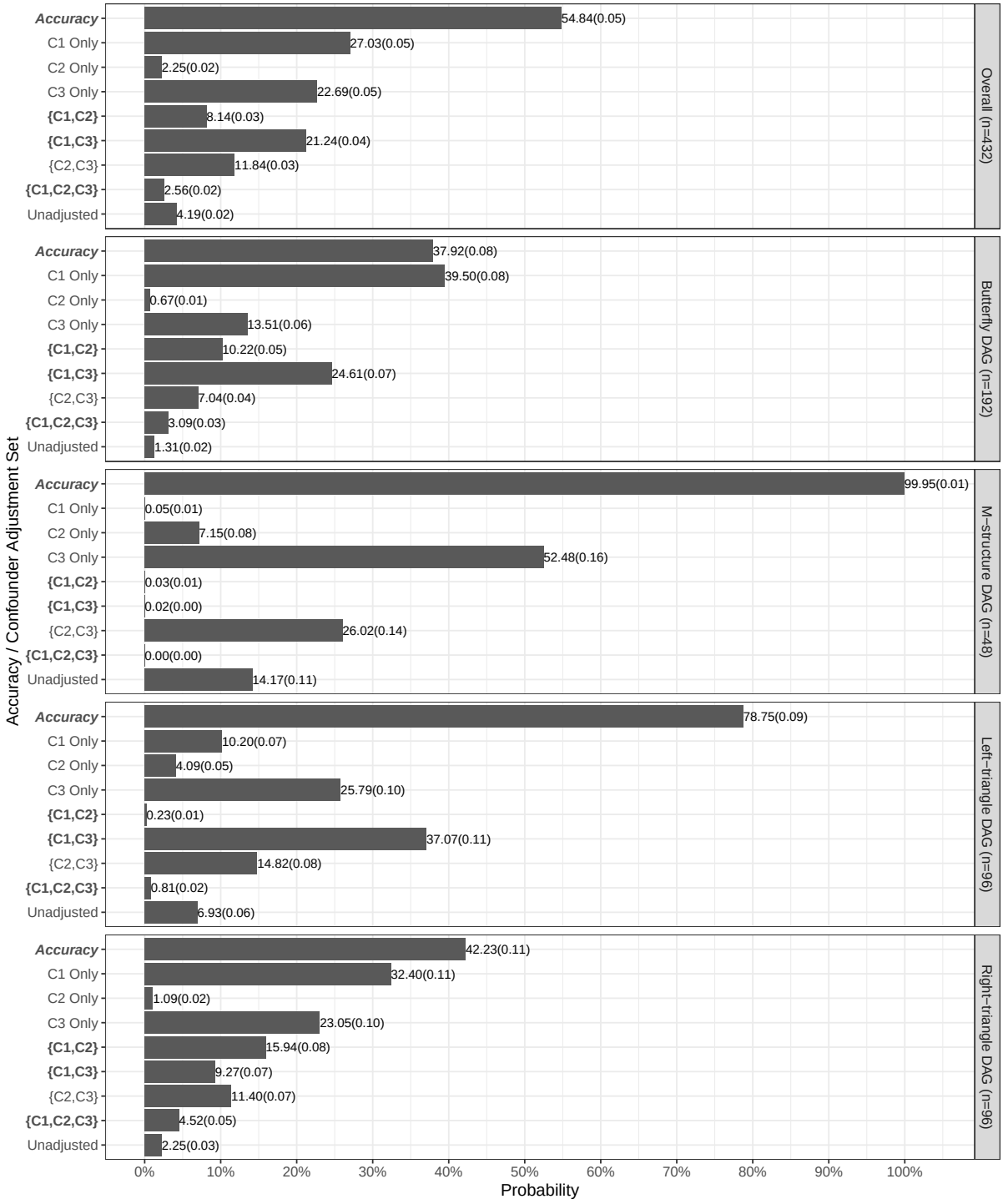
Overall, both the default and the tuned MMHC algorithms yielded low accuracy in selecting appropriate confounder adjustment sets. Figure 3 shows that the accuracy of MMHC is highest in M-structure DAGs, followed by left-triangle, right-triangle and butterfly DAGs. These results are also provided in tabular format in Appendix A.3. The difference in accuracy between the four types of DAGs is related to the number of admissible confounder adjustment options and the probability of C_1, C_2 and C_3 being selected. For example, the butterfly structures have only three admissible adjustment sets that block all backdoor paths, whereas the left- or right-triangle structures each have five admissible adjustment sets that block all backdoor paths.

Figure 4 shows the probability that a confounder is selected by MMHC. In the butterfly DAG, the probability of selecting C_1 is much higher than C_2 and C_3, resulting in an insufficient adjustment set (also see Figure 3: 40% in selecting C_1 only). Besides that, C_3 is more likely to be selected than C_2 in all types of DAG. This explains the higher accuracy of the MMHC algorithms in left-triangle DAGs than in right-triangle DAGs, because any adjustment sets containing C_3 are admissible for left-triangle DAGs while sets containing C_2 for are admissible right-triangle DAGs. The nuance variables B were almost never selected by the MMHC algorithms.

The tuned MMHC performs slightly better than the default one (overall: 54.8% vs

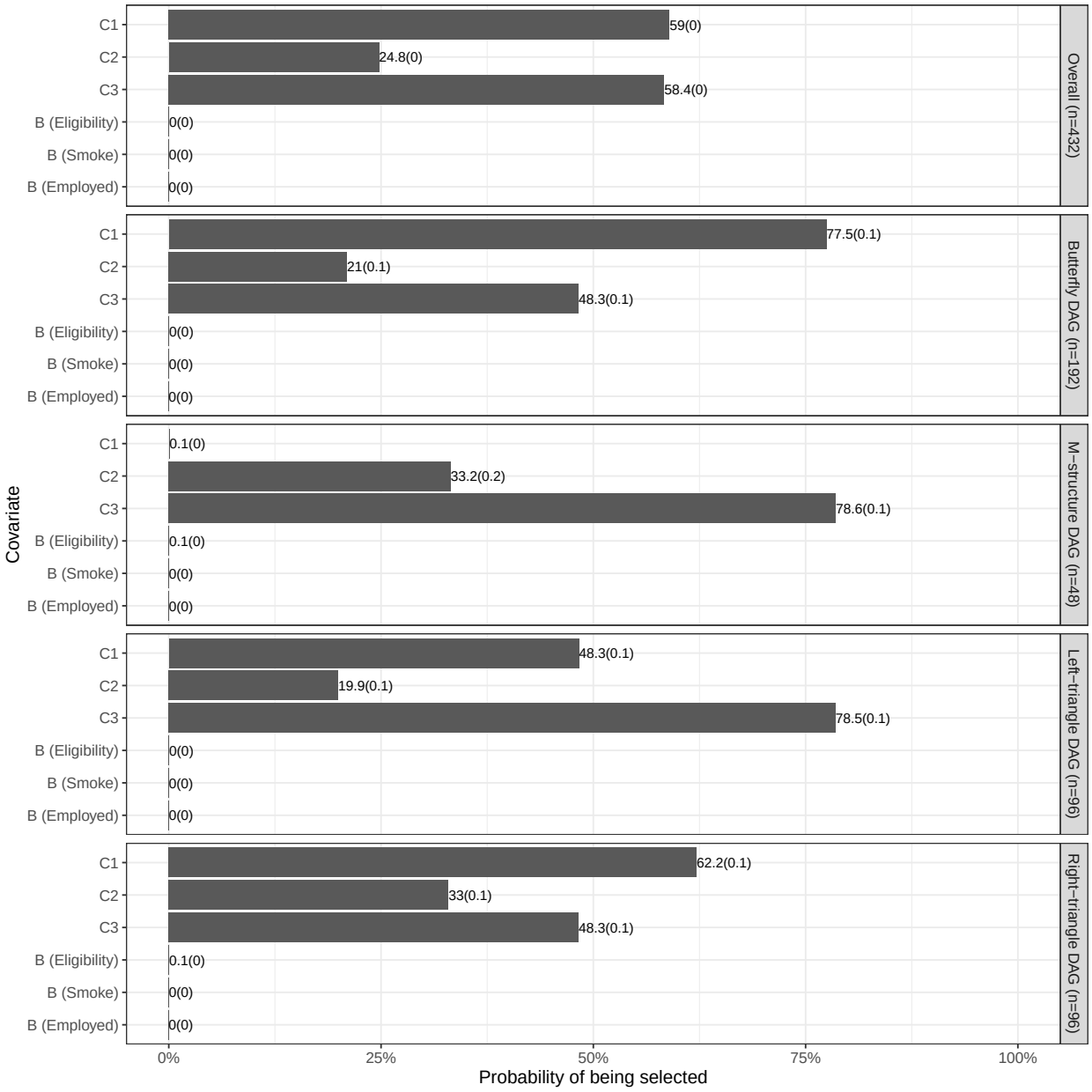
52.8% in Figure 11), since the tuned MMHC is more likely to select confounder adjustment sets containing C_2 and less likely to yield an empty adjustment set (Figure 11 & 12). As such, tuning MMHC only improves the accuracy in right-triangle (42.2% vs. 38.7%) and butterfly DAGs (37.9% vs 35.1%).

Figure 3: Accuracy of the tuned MMHC algorithm in selecting correct confounder adjustment set



- Numbers within parentheses are SE
- **Bold** sets are common admissible adjustment for all 432 DAGs.

Figure 4: Probability of covariates selected by the tuned MMHC, stratified by DAG type



- Numbers within parentheses are SE

2.3.2 Absolute bias and MSE of ATE estimation using manual and automated adjustment set selection

Table 4 presents medians and interquartiles of absolute bias and MSE in estimating ATE. With manual confounder selection, the median absolute biases are small and similar in all of our scenarios (median range$= 0.9 - 5.8 \cdot 10^{-4}$). Automated confounder selection using either MMHC algorithm yields slightly higher bias than all manual selection scenarios except the unadjusted estimates. The tuned MMHC algorithm has better performance than the default, as expected. Median and interquartiles of MSE are similar among all scenarios regardless of confounder adjustment scenarios.

Figure 13-14 in the Appendix A.5 show the distributions of the absolute bias and MSE among 432 DAGs. Within a specific confounder selection scenario and one type of DAG, g-computation, IPW, AIPW and TMLE have similar distributions of absolute bias (Figure 13). In the manual confounder selection settings, there is no visual difference in the absolute bias distributions between scenarios adjusting for all causes and for all covariates. Interestingly, although an empty adjustment set is admissible for an M-structure DAG, the distributions of absolute bias for unadjusted estimates are similar to those adjusting for the collider only. This is likely due to the negligible impact of collider bias in a range of scenarios.[32] Compared to manual selection, automated selection generally yields higher absolute biases only in the butterfly-structure DAG but no visible difference in other types of DAGs. In Figure 14, g-computation has lower MSE when adjusting for all causes or covariates in the manual selection setting, compared to all other scenarios or estimators.

Table 4: Absolute bias and mean squared error (MSE) of average treatment effect from different confounder adjustment sets by manual and automated selection methods

	Abs. Bias ($\cdot 10^{-4}$)	Relative Abs. Bias	MSE ($\cdot 10^{-4}$)
CovSet	Median (Q1, Q3)	Median (Q1, Q3)	Median (Q1, Q3)
Manual Confounder Selection			
All Causes (with Collider)[✓]	0.90 (0.44, 1.51)	(Reference)	1.09 (1.06, 1.12)
All Covariates[✓]	0.91 (0.42, 1.49)	1.01 (0.97, 0.99)	1.11 (1.08, 1.14)
Collider Only	1.23 (0.63, 2.24)	1.37 (1.43, 1.49)	1.10 (1.07, 1.13)
Empty Set / Unadjusted	5.82 (2.69, 17.39)	6.48 (6.17, 11.53)	1.16 (1.11, 1.40)
Automated Confounder Selection: MMHC default			
All Causes	1.71 (0.78, 3.11)	1.90 (1.80, 2.06)	1.11 (1.08, 1.15)
Automated Confounder Selection: MMHC tuned			
All Causes	1.58 (0.74, 2.83)	1.76 (1.69, 1.87)	1.11 (1.08, 1.15)

[✓] Admissible adjustment set that blocks the backdoor path from A to Y for all 432 DAGs
[1] Median (Q1, Q3) of absolute bias relative to adjusting for all causes with manual confounder selection
[2] Accuracy of correct confounder adjustment set is evaluated with causes of A and/or Y

2.4 Discussion

We evaluated the performance of causal discovery algorithms in identifying relevant confounders using plasmode simulations with data generated from an actual epidemiologic study. Our results suggest current causal discovery algorithms underperform in selecting confounder adjustment sets using the data with small to moderate effect sizes and sample sizes. While the data generating mechanisms we explored were complex, they arguably align with many scenarios of epidemiologic interest.[72, 36] As a result, we

found that using automated confounder selection methods as a step in the process of estimating causal effects would yield higher bias than a more traditional approach involving manual confounder selection with the knowledge of true data generating mechanisms.

Causal discovery algorithms have been used and evaluated in other fields.[71, 78, 82, 70, 83] This prior work has shown that causal discovery algorithms hold some promise in correctly identifying underlying DAGs with minimal background knowledge.[71, 83] However, thus far, most evaluations have relied on data that are highly dissimilar to those encountered in epidemiology. This includes datasets simulated to resemble gene expression data, where all relevant variables follow a standard normal distribution, or where variables are all discrete, and in settings where effect sizes are large.[78, 84, 85, 86, 87, 88]

Unfortunately, most epidemiologic studies rely on data that consist of a mix of continuous, discrete, and ordinal variables, potentially with small effect sizes, and small to moderate sample sizes. These complications have implications for how well causal discovery algorithms can perform, as we have shown. For example, mixing ordinal and discrete variables can influence average effect sizes. In our simulations, despite having the same conditional associations, the marginal effect size of C_3 is larger than C_2, because C_3 (hsCRP) has three levels while C_2 (number of prior pregnancy losses) has only two. As such, C_3 is more likely to be selected than C_2 by a causal discovery algorithm, explaining in part the low accuracy of the MMHC algorithms.

The MMHC algorithm has shown good performance in simpler graphical structures with high-dimensional data.[70, 79] For example, random graphs, a simple graph structure whose edges are generated based on a probability distribution,[89] are often used for evaluating causal discovery algorithms.[90, 88] However, the underlying causal struc-

tures of epidemiologic data are more complex than those evaluated DAGs.[60] As such, our M-/Butterfly-structure DAG presents a complex but realistic challenge for these methods.[72, 36] Although we had incorporated with a reasonable amount of background knowledge into the MMHC algorithms (e.g., forcing the directions), we still observed a low accuracy of causal discovery methods. Our results imply that the performance would be worse if the automated confounder selection is purely data-adaptive without imposing any background knowledge.

In our study, using causal discovery to select confounders yielded higher absolute bias in ATE estimation than adjusting for all covariates in the dataset. This result is applicable in low-dimensional settings since our simulated dataset only contained 11 covariates. However, in high-dimensional setting, adjusting for all covariates may not be feasible for several reasons, including the curse of dimensionality,[91, 92] or the inclusion of variables that introduce or amplify bias.[63, 93] Hence, causal discovery may be useful for pre-screening confounders in high-dimensional data.[62] Notably, before using these algorithms, investigators may need to preprocess the data (e.g., discretization),[80, 87] use sophisticated tuning methods (e.g., imposing structural priors),[66, 94] and consider the assumptions and limitations of different causal discovery methods (e.g., parametric assumptions used for hypothesis testing in Appendix A.2, causal faithfulness condition, inability to distinguish DAGs in the same Markov equivalence class).[95, 96]

In conclusion, while causal discovery algorithms hold some promise, epidemiologists should be aware of their limitations in accurately selecting sufficient confounder adjustment sets. Given the uncertainty of true data generating mechanisms in observational studies, use of sophisticated causal discovery based variable selection algorithms should be accompanied with an appropriate degree of caution. Future development of causal

discovery methods is needed for epidemiologic studies, such as better handling a mixed type of data, incorporating statistical models that perform well with small effect sizes, and adapting for complex causal structures.

3.0 AIPW: An R Package for Augmented Inverse Probability Weighted Estimation of Average Causal Effects

3.1 Introduction

Machine learning methods are increasingly being used to estimate cause-effect relations. Numerous examples exist, including using random forests, gradient boosting, or a combination of learners (e.g., stacking) for propensity score weighting, stratification, or matching, or via marginal standardization with a regression based estimator.[97, 98, 99, 100, 101, 53] However, there is a growing body of theoretical and simulation evidence suggesting that, without some form of statistical bias correction, using machine learning methods to estimate causal effects can result in high bias, high mean squared error (MSE), and less than nominal confidence interval (CI) coverage. [46, 47, 102, 49, 50]

In contrast, doubly robust estimators possess a statistical bias correction property,[103] and are thus less susceptible to problems with bias, MSE, and CI coverage when machine learning methods are used. Hence, when estimating causal effects with machine learning methods, doubly robust estimators, such as targeted maximum likelihood estimation (TMLE) or augmented inverse probability weighting (AIPW), should be used. [49, 104, 105, 102, 50] Several software programs that implement doubly robust estimators are currently available in a number of different programming languages, including SAS,[54] Stata,[55] R,[75, 106, 56, 107, 108] python,[109] and MATLAB.[110] However, only a handful of them enable use of machine learning methods.[75, 106, 107] Additionally, most share important limitations known to either affect the performance of doubly robust estimation, or lower their relevance to epidemiologists. Most importantly,

these limitations include the inability to implement sample splitting or cross-fitting, and the estimation of effects on a single scale of measurement (e.g., additive effects). To address these limitations, we developed the **AIPW** package, which implements augmented inverse probability weighting [51] for a binary exposure in the R programming environment.[81] Compared to other packages for implementing doubly robust estimators via machine learning methods, the **AIPW** package:

1. allows different covariate sets to be specified for the exposure and the outcome models, which may be important when analyzing data from randomized trials;
2. obtains appropriate standard errors for estimates of the average treatment effect by implementing k-fold cross-fitting;
3. relies on a user-friendly parallel processing framework for computationally heavy tasks;
4. enables estimation directly from the fitted objects from existing doubly robust implementations (e.g., tmle[75], or tmle3[106]) in the R programming language.

In this paper, we illustrate the AIPW estimator, and how to use it in our package. Additionally, we highlight the differences between various software implementations of these estimators in the R programming language, including **AIPW**, **CausalGAM**,[56] **npcausal**,[107] **tmle** [75] and **tmle3**.[106]

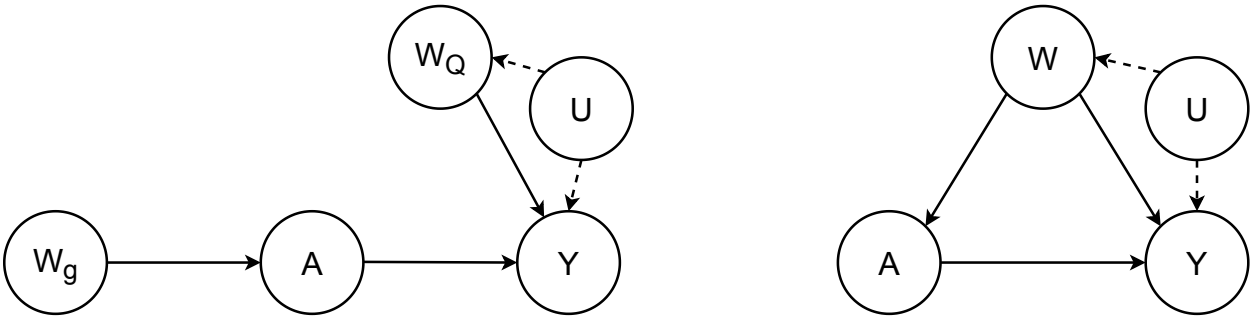
3.2 Methods

3.2.1 Motivation and data generating mechanisms

Here we outline the datasets motivating our illustration of augmented inverse probability weighting, and the use of the **AIPW** package. We rely on the Effects of Aspirin in Gestation and Reproduction (EAGeR) study, a multicenter randomized trial of the effect of daily low-dose aspirin on pregnancy outcomes in women at high risk of miscarriage. The trial recruited 1,228 women aged 18-40 years attempting to become pregnant. Details on the EAGeR trial and data are described elsewhere.[12, 13, 14, 15]

We simulate two different datasets from EAGeR to illustrate the use of the **AIPW** package. We use a simulation approach because: (i) the actual data are not publicly available; and (ii) true exposure effects are known in simulation settings. Data are generated based on the causal relations depicted in Figure 5.

Figure 5: Causal diagrams for a randomized controlled trial and an observational study



(a) Randomized Controlled Trial (b) Observational Study

Figure 5a illustrates a data generating mechanism for a randomized trial in which the treatment A is assigned conditionally based on a measured covariate W_g. For example,

in a study designed to explore the impact of aspirin on pregnancy outcomes in women with previous pregnancy losses, one may decide to randomize to aspirin versus placebo 1:1 for women with only one prior pregnancy loss, but elect to randomize 3:1 for women with more than one prior pregnancy loss. Similarly, Figure 5b illustrates a simple causal diagram for an observational study of the relation between an exposure A (e.g., whether a given woman took aspirin during the study's follow-up), an outcome of interest Y (e.g., an indicator of whether live birth occurred during follow-up), and a set of confounders of the exposure-outcome relation W.

To construct datasets governed by the data generating mechanisms in Figure 5, we sampled (with replacement) baseline covariates from the EAGeR data. For the simulated RCT (N=1,228, Figure 5a), A denotes the binary treatment assignment, Y is the binary outcome, W_g represents the covariate that affects the treatment assignment, which in our case was deemed to be the eligibility stratum indicator, sampled with replacement from the EAGeR trial. Similarly, W_Q is a set of baseline prognostic covariates, which were also sampled with replacement from the EAGeR trial, and included the number of prior pregnancy losses, age, the months of conception attempts prior to randomization, BMI and mean arterial blood pressure (denoted as $W_{1...5}$, respectively). Our simulated treatment A was generated such that $P(A = 1|W_g = 1) = 0.75$ and $P(A = 1|W_g = 0) = 0.25$. The outcome Y was simulated from a logistic regression model defined as:

$$\text{logit}[P(Y = 1 \mid A, W_Q)] = 2.20 + 0.56A + 0.05W_1 - 0.01W_2 - 0.08W_3 - 0.03W_4 - 0.01W_5$$

The above model defines the treatment effect via a conditional OR of 1.75. In our simulated setting, this yielded true marginal effects of 0.13, 1.29 and 1.71 on the risk difference, risk ratio, and odds ratio scales, respectively (Table 5, row 1). We used the correctly

specified parametric regression model in a sample of 1 million observations to obtain the estimate of the true effects to serve as our parameters of the true causal effect parameter values.

For the simulated observational study governed by the data generating mechanism in Figure 5b, A, Y and W denote a binary exposure, a binary outcome, and a set of binary, categorical, and continuous confounders (i.e., the aforementioned W_g and $W_{1...5}$), respectively. The propensity score model used to generate A was defined as:

$$\text{logit}[P(A = 1 \mid W)] = -0.29 + 0.56W_g - 0.23W_1 + 0.01W_2 + 0.02W_3 - 0.02W_4 + 0.01W_5$$

Similarly, the outcome Y was simulated from an outcome model defined as:

$$\text{logit}[P(Y = 1 \mid A, W)] = 2.03 + 0.56A - 0.37W_g + 0.30W_1 - 0.01W_2 - 0.08W_3 - 0.05W_4 - 0.01W_5$$

such that the true conditional OR for the exposure-outcome relation was 1.75. This yielded true marginal effects of 0.13, 1.36, and 1.70 on the risk difference, risk ratio, and odds ratio scales, respectively, which were again obtained using the approach described above.

Realizations of both of these datasets are included in the AIPW package, and can be obtained using the `data(eager_sim_rct)` and `data(eager_sim_obs)` function.

3.2.2 Basic implementation of AIPW

The **AIPW** package was developed to estimate treatment effects of a binary exposure. Such effects include average treatment effects (ATE) commonly targeted in observational studies, which include intention-to-treat (ITT) effects when a randomization indicator is available. These effects can be defined on the risk difference (RD), risk ratio (RR) and

odds ratio (OR) scales as:[111]

$$
\begin{aligned}
RD &= E(Y^1 = 1) - E(Y^0 = 1) \\
RR &= \frac{E(Y^1 = 1)}{E(Y^0 = 1)} \\
OR &= \frac{E(Y^1 = 1)}{1 - E(Y^1 = 1)} \Big/ \frac{E(Y^0 = 1)}{1 - E(Y^0 = 1)}
\end{aligned}
$$

where Y^1 and Y^0 denote the potential outcomes that would be observed if the exposure was set to 1 and 0, respectively.

Under consistency, exchangeability, positivity, and no interference, the average of potential outcomes that would be observed under $A = a$ are identified as the average of estimated outcomes, that is: $E(Y^a) = E[E(Y \mid A = a, W)]$, which for simplicity we denote as $\psi(a)$. Several estimators can be constructed by combining predictions from the propensity score model with predictions from the outcome model. These predictions can be obtained from parametric regression, such as logistic regression. However, machine learning methods can also be used when these predictions are combined via a doubly robust estimator such as AIPW. This is because double robustness can yield estimators with low bias and valid standard errors, even when the propensity score and outcome model estimators have high bias and no generally valid method for obtaining standard errors.[46, 47, 102, 49, 50]

Under the data generating mechanism depicted in Figure 5a, the propensity score predictions should be obtained conditional on W_g (i.e., $\hat{P}_i(A = 1|W_g)$), which could be used for constructing an inverse probability weighting estimator, such as: [74, 112]

$$
\hat{\psi}_{IPW}(a) = \frac{1}{n} \sum_{i=1}^{n} \frac{I(A_i = a)}{\hat{P}(A = a \mid W_{g,i})} \cdot Y_i \tag{4}
$$

where $a \in \{0, 1\}$ and i represents i^{th} observation. For improved performance, the es-

timated propensity scores can be truncated, which the **AIPW** package implements by default at the 2.5% percentile.[113]

Alternatively, outcome model predictions $\hat{P}(Y = 1 \mid A, W_Q)$ can be used to construct a g computation estimator, defined as [114, 74]

$$\hat{\psi}_{gComp}(a) = \frac{1}{n}\sum_{i=1}^{n} \hat{P}(Y = 1 \mid A \coloneqq a, W_{Q,i}) \tag{5}$$

where the $\coloneqq$ symbol denotes that we set each individual's value for A in the sample to the argument's value a. This equation represents the average of predictions from the outcome model by setting $A = a$ over each confounder level.

When the propensity score model or the outcome model are used alone to estimate average treatment effects, they must in general be built from correct parametric models. In contrast, one can use both the propensity score and the outcome models together in an AIPW estimator as follows:[114, 51, 103, 56, 115]

$$\hat{\psi}(a)_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(A_i = a)}{\hat{P}(A = a|W_{g,i})}[Y_i - \hat{P}(Y = 1|A_i, W_{Q,i})] + \hat{P}(Y = 1|A \coloneqq t, W_{Q,i})\right\} \tag{6}$$

A TMLE estimator of the same quantities can also be constructed using alternative techniques.[75, 105]

As with the TMLE estimator, missing outcome data can be accounted for with the AIPW package if the covariate set W (i.e., both W_Q and W_g) enables one to assume outcomes are missing at random conditional on W (Appendix B.1).[75, 116]

As long as either the outcome or the exposure model is correctly specified, consistent estimates of the mean potential outcome can be obtained; i.e., the doubly robust property of AIPW.[117] Additionally, because of certain statistical properties of doubly robust estimators,[49] one can use machine learning methods to quantify the exposure and

the outcome models while minimizing the slow convergence rates (i.e., large MSE) and overfitting problems that typically characterize use of machine learning methods with sample-splitting or cross-fitting.[49, 50] Appendix B.2 describes the implementation of cross-fitting used in the **AIPW** package, as well as a general description of the relation between sample splitting and cross-fitting.

Standard errors (SE) for the AIPW on the RD scale can be constructed by taking the standard deviation of the estimated efficient influence function evaluated at each observation.[118] Similarly, standard error estimates for the estimated RR and OR can be constructed using the delta method. All derivations are provided in Appendix B.3.

3.2.3 Package implementation

The **AIPW** package can easily be used to obtain ATE estimates on the RD, RR, and OR scales in several different ways. Using the simulated RCT data provided in the package, Figure 6 provides some example code that could be used to obtain the results presented in Table 5, row 2.

Figure 6: Example code that can be used to implement an augmented inverse probability weighted estimator via the **AIPW** package using the simulated RCT data available in the package.

```
library(AIPW)
library(SuperLearner)
set.seed(1234)
#load simulated dataset (RCT)
data(eager_sim_rct)
#Specify SuperLearner libraries
sl.lib = c("SL.gam","SL.earth","SL.ranger","SL.xgboost")
#Create a vector of covariates
Cov = c("loss_num","age", "time_try_pregnant","BMI","meanAP")
#create a new AIPW object called AIPW_SL
AIPW_SL <- AIPW$new(Y = eager_sim_rct$sim_Y,
                    A = eager_sim_rct$sim_Tx,
                    W.g = eager_sim_rct$eligibility,
                    W.Q = subset(eager_sim_rct,select=Cov), #covariates
                    Q.SL.library = sl.lib, #outcome model
                    g.SL.library = sl.lib, #exposure model
                    k_split = 10, #num of folds for cross-fitting
                    verbose=TRUE)
#fit the data stored in the AIPW_SL object
AIPW_SL$fit()
#summarise the results using truncated propensity scores
AIPW_SL$summary(g.bound = 0.025)
```

The **AIPW** package was developed with the object-oriented programming design via the R6 class.[119, 120] Similar to TMLE, the AIPW function can employ the Super Learner stacking algorithm.[121, 122] In the example code in Figure 6, we combine four learners via stacking, including generalized additive model (GAM in **gam** package),[123] multivariate adaptive regression splines (**earth**),[124] random forests (**ranger**)[125] and extreme gradient boosting (**xgboost**)[126] to fit the propensity score and outcome models. Additionally, the **AIPW** function enables k-fold cross-fitting, which can provide more accurate standard error estimates when machine learning methods are used.[127, 49] Users must specify the $k\_split \geq 2$ argument to enable cross-fitting for the **AIPW**. This `AIPW_SL` object

is then fitted with the stored arguments using `fit()` as depicted in line 20 of Figure 6, and the results are summarised using the `summary()` function (line 22). The propensity score can be truncated using the $g.bound$ argument in `summary()`: propensity scores lower than $g.bound$ or higher than $1 - g.bound$ are set to $g.bound$ or $1 - g.bound$, respectively. For comparison, results from corresponding software implementations are also provided in Table 5.

Full details on using the **AIPW** are available from the Comprehensive R Archive Network at `http://CRAN.R-project.org/package=AIPW` or our Github repository at `https://github.com/yqzhong7/AIPW`. This includes details on a range of scenarios that may be encountered with data in randomized trials or observational studies, as well as options in the **AIPW** package that can be used to tailor analyses. In addition, methods for providing average treatment effects among the treated (ATT) along with their SE are described online and in the package help documentation.[128]

3.2.4 Performance evaluation via a simulation study

To evaluate the performance of our **AIPW** package, and compare it to existing implementations of double-robust estimators, we conducted a simulation study in observational study data. A sample of $n = 200$ from the observational data generating mechanism (Figure 5b) is provided with the **AIPW** package. We use this data generating mechanism to evaluate and compare **AIPW** and other doubly robust implementations in the R programming language (i.e., **CausalGAM**, **npcausal**, **tmle**, **tmle3**).[56, 107, 75, 106] Two thousand Monte Carlo simulations, each with a sample size of 200 observations, were conducted. Because **CausalGAM** does not support estimating effects on the multiplicative scale, we only evaluated the performance for the RD scale. Performance was eval-

uated via estimated bias [$E(\widehat{RD}) - RD_{true}$] and MSE [$E[(\widehat{RD} - RD_{true})^2]$] for the point estimates, as well as mean 95% CI width [$E(\widehat{RD}_{upper} - \widehat{RD}_{lower})$] and 95% CI coverage [$P(\widehat{RD}_{lower} < RD_{true} < \widehat{RD}_{upper})$] for the asymptotic standard errors.[129] We also provide information on mean runtime per Monte Carlo run (in seconds; sequentially, without parallel processing).

To explore the performance of different estimators, five sets of analyses were performed. First, the true outcome and propensity score models (GLMs) were used to estimate the RD in all five packages along with g-computation (via the true outcome model) and stabilized inverse probability weighting (IPW, via the true propensity score model). Second, only generalized additive models (GAMs) (**gam**) were used to estimate the RD without cross-fitting in each of the five packages implementing doubly robust estimators. Third, GAMs were used with 10-fold cross-fitting for **AIPW**, **npcausal**, **tmle** and **tmle3** packages, the only four packages that enable implementation of cross-fitting. Fourth, we used the Super Learner to stack **gam**, **earth**, **ranger** and **xgboost** into one meta-algorithm for RD estimation in **AIPW**, **npcausal**, **tmle** and **tmle3** without cross-fitting.[121, 122, 130, 131] Because **CausalGAM** only supports GAMs, we could not evaluate this package with the stacked meta-learner. Lastly, we repeated the latter AIPW and TMLE analyses, but this time using 10-fold cross-fitting, using the **AIPW**, **npcausal**, **tmle** and **tmle3** packages. Simulations were conducted in R (version 3.6.2) and details about the models used for estimation (e.g. tuning parameters) are provided in the Github repository at `https://github.com/yqzhong7/AIPW_Simulation/blob/main/AIPW_simulation.md`.

3.3 Results

Table 5 presents the ATE estimates from the four doubly robust packages in the example RCT data provided with the package. When estimated via the **AIPW** package, we obtained a risk difference of RD_{AIPW}=0.136 (95%CI 0.070, 0.201) for the average treatment effect if all subjects were treated versus untreated. Similarly, the corresponding risk ratio and odds ratio obtained from the **AIPW** package were RR_{AIPW}=1.305 (95%CI 1.143, 1.490) and OR_{AIPW}=1.727 (95%CI 1.323, 2.253). Additionally, despite the differences in implementation and estimation, the other packages yielded estimates that were consistent with those obtained from **AIPW**. Estimates from all packages were close to the true estimates.

Performance results from our simulations are shown in Table 6. In general, among 2000 simulated observational datasets, each with a sample size of 200, there was no substantive difference in the bias and MSE between any of the packages used. As expected, the bias using GLMs and GAMs were similar, but were generally lower than the bias using Super Learner. Among packages using GAMs, we observed that the **CausalGAM** yielded a bias about twice as the **AIPW**, **npcausal**, **tmle** and **tmle3**. Among the packages enabling Super Learner without cross-fitting, bias of the **AIPW** and **tmle** were about twice as **npcausal** and **tmle3**. In terms of 95% confidence intervals, coverage was less than nominal [i.e., $P(\widehat{RD}_{lower} < RD_{true} < \widehat{RD}_{upper}) < 95\%$] without cross-fitting except when correct parametric models used, while the coverage improved to nominal when cross-fitting was enabled. Notably, cross-fitting in our setting largely improved the performance of the **AIPW** package, especially using Super Learner—its bias decreased from -0.009 to -0.002 and CI coverage increased from 93.0% to 95.6%—which are comparable

to its performance using the true GLMs (Bias = -0.002 and CI Coverage = 94.8%).

Figure 7 shows the pairwise comparisons of the ATE estimates from the simulation results using GLMs in Table 6. Panels on the diagonal are the distributions of estimates, and the lower triangular area includes pairwise scatter plots of all estimates. In the scatter plot panels vertical and horizontal lines both depict the $RD_{true} = 0.13$. Estimates near the intersection of the true RD lines are less biased from both methods compared in the scatter plot. Interestingly, the estimates are highly correlated between the singly robust estimators (Pearson's correlation between g-computation and IPW =0.99) and among doubly robust estimators (Pearson's correlations ≥ 0.97), respectively; however, the correlations are only moderate between singly and doubly robust estimators (Pearson's correlation = 0.44). Similarly, Appendix B.4 shows the pairwise comparisons of the ATE estimates using GAMs and Super Learner in Table 7; all packages also yielded highly correlated estimates despite the different estimation methods.

3.4 Discussion

In this paper, we presented a new R implementation of the augmented inverse probability weighted (AIPW) estimator, via the **AIPW** package. This package provides a flexible implementation of the AIPW estimator via stacking (e.g., super learner with parametric and machine learning algorithms). Designed for randomized trials and observational studies, the **AIPW** package can provide average causal effect estimates for a binary exposure on the risk difference, risk ratio, and odds ratio scales, as well as support various features such as cross-fitting, parallel processing, and allowing different covariate sets for the exposure and the outcome models.

For convenience, we summarised the key functionality of the **AIPW** package and its comparisons with **CausalGAM**, **npcausal**, **tmle** and **tmle3** in Table 7. Comparing the two packages implementing augmented inverse probability weighting, the **AIPW** package is more flexible than the **CausalGAM** because it supports estimations in multiplicative scales, models using stacking machine learning algorithms via **SuperLearner**[130] or **sl3**[131], and cross-fitting. Compared to **tmle** and **tmle3**, the **AIPW** package holds similar features, and additionally, it supports using the fitted **tmle** and **tmle3** object as input for AIPW estimation.

Indeed, while often used in observational data, doubly robust estimators can be important when analyzing data from randomized trials; in fact, they can be asymptotically efficient under essentially no assumptions. In such a setting, researchers may often wish to adjust for covariates to increase the efficiency of the unconditional intention-to-treat effect.[132, 133, 134, 58] However, when adjusting for covariates, one may inadvertently introduce misspecification biases, thus detracting from one of the major benefits of randomization.[133, 134] Notably, doubly robust estimators can avoid such biases for randomized trials, because the data generating mechanism for treatment allocations (i.e., randomization stratum) is known by investigators.

Adjusting for covariates in an RCT via double robust estimation requires considering different covariate sets for the propensity score and outcome models. For instance, covariates that were not used to assign treatment need not generally be included in the exposure model, even though they might be included in the outcome model. The **AIPW** package easily allows specifying different covariate sets for the outcome and the exposure models, and can thus be used for doubly-robust estimation in randomized trials. In addition, the **AIPW** package enables model specification using machine learning methods,

which can avoid the strict assumptions imposed by parametric models.

With the observational data, our simulation study shows comparable performance of the **AIPW** package relative to other packages. Indeed, excellent performance was observed even with a relatively small sample size (n=200). Performance would be expected to improve as the sample size increases.[49]

Cross-fitting yielded major improvements in bias and confidence interval coverage of doubly robust methods in our simulation study, in line with a growing body of literature.[46, 47, 102, 49, 50] Intuitively, sample splitting or cross-fitting can be used to mitigate overfitting. If cross-fitting is not used, the same data would be used twice for two different tasks—once for estimating nuisance quantities (i.e., propensity scores and outcome model predictions), and once for averaging over them to form the estimator.[127, 47] Mathematically, cross-fitting (along with consistency of nuisance estimators, at any rate) ensures a so-called empirical process term is asymptotically negligible—without sample splitting one would need to rely on unverifiable assumptions about the true model that may not hold with high dimensional data.[135] Hence, complex machine learning methods should be accompanied by sample splitting or cross-fitting for effect estimation.

Many machine learning methods, along with cross-validation, sample splitting or cross-fitting procedures, often rely on pseudo-random number generators to complete the estimation procedure. With such procedures, reproducibility can be attained by setting "seeds" that determine the exact settings in which the pseudo-random number generators operate. Unfortunately, this can make the results from a given study highly dependent on the value of the selected seed, particularly when cross-fitting is used. Several options are available that reduce the extent to which results depend upon a selected seed value. These include using a higher number of folds for cross-fitting, repeating the cross-fitting

procedure iteratively in a given dataset,[47, 136] or, if one is willing to make unverifiable assumptions (i.e., Donsker condition), avoiding cross-fitting entirely.[135]

At present, the **AIPW** package relies on a single application of cross-fitting, which may result in seed dependence. Future versions of the package will include options for an iterative cross-fitting procedure. However, users concerned about seed dependence in the current package could select a large number of cross-fitting folds to mitigate this potential issue.

Theoretically, AIPW and TMLE estimators are asymptotically equivalent. Differences between the two arise only due to finite sample differences. These relationships are presented in Figure 7, and Appendix B.5 & B.6 with a sample size of 200 from 2000 Monte Carlo samples. It also provides a degree of validation for our **AIPW** package by comparing it to existing, well-known, doubly robust R programs.

Our implementation of AIPW estimation is based on a particularly well-studied estimator.[51, 103, 128] However, it is important to note that there are several different variations of the AIPW estimator distinct from the one we use. Some of these are known to perform better in certain settings, such as when there are potential near positivity violations.[116, 137] Our use of propensity score truncation does alleviate some of the concerns raised by such positivity violations. Yet researchers should be aware of the existence of alternative AIPW estimation methods.

Future planned implementations for the **AIPW** package include supporting categorical exposures by incorporating missing data mechanisms,[116, 75] and an iterative cross-fitting procedure.[47, 136] Runtime of the **AIPW** package depends on the algorithms included in the stacked learner and the implementation of stacking. Our preliminary (and unvalidated) findings suggest that the **sl3** package is faster than the **SuperLearner**.[131]

For convenience, we find that using **SuperLearner** for small jobs and **sl3** for more complex models tends to optimize run time.[131] Furthermore, to optimize run time, we have enabled use of parallel processing packages available in R. Given the **AIPW** package is hosted on GitHub, future maintenance (e.g., bug reporting) can be requested on GitHub issues.

Altogether, doubly robust estimators are a powerful tool to investigate cause-effect relations with machine learning methods. The novel **AIPW** package addresses the limitations of existing programs implementing doubly robust estimators and facilitates epidemiologists to conduct causal inference with flexible machine learning methods.

3.5 Tables and Figures

Table 5: Estimated average treatment effects of a simulated randomized controlled trial based on EAGeR

	RD				RR				OR			
Package	Est.	SE	LCL	UCL	Est.	SE	LCL	UCL	Est.	SE	LCL	UCL
True Est.	0.132	-	-	-	1.285	-	-	-	1.708	-	-	-
AIPW	0.136	0.033	0.070	0.201	1.305	0.068	1.143	1.490	1.727	0.136	1.323	2.253
CausalGAM	0.134	0.033	0.070	0.198	-	-	-	-	-	-	-	-
npcausal	0.133	0.035	0.065	0.201	-	-	-	-	-	-	-	-
tmle	0.135	0.026	0.083	0.186	1.306	0.054	1.176	1.451	1.719	0.107	1.394	2.121
tmle3	0.138	0.034	0.071	0.205	1.310	0.070	1.141	1.503	1.764	0.140	1.339	2.323

[1] SE are asymptotic estimation (by delta method)
[2] SuperLearner was used for AIPW, npcausal and tmle, and sl3 for tmle3. Algorithms include GAM, earth, ranger and xgboost.
[3] 10-fold cross-fitting was use for AIPW, npcausal, tmle and tmle3. (tmle only support cross-fitting in the outcome model.)
[4] Three different estimations were done for tmle3 since it can only output one type of estimand per estimation
[5] The estimates of true causal effect parameter values were generated by the correctly specified parametric regression model with a sample size of 1 million (Figure 1a)

Table 6: Performance of the AIPW package in estimating the average treatment effect (risk difference) in a simulated observational study based on EAGeR

Package/Method	Bias (SE)	MSE	MeanCIwidth	Coverage (SE)	MeanRuntimeSec
True Model: GLM + No sample splitting					
gComp	-0.002 (0.002)	0.005	0.271	94.8% (0.5%)	1.82
IPW	-0.002 (0.002)	0.005	0.280	95.8% (0.4%)	0.01
AIPW	-0.002 (0.002)	0.005	0.268	94.8% (0.5%)	0.36
CausalGAM	-0.003 (0.002)	0.005	0.267	94.8% (0.5%)	0.07
npcausal	-0.002 (0.002)	0.005	0.267	94.6% (0.5%)	0.24
tmle	-0.002 (0.002)	0.005	0.261	94.4% (0.5%)	0.29
tmle3	-0.002 (0.002)	0.005	0.268	94.8% (0.5%)	0.31
GAMs + No sample splitting					
AIPW	-0.002 (0.002)	0.005	0.261	93.8% (0.5%)	1.16
CausalGAM	-0.004 (0.002)	0.005	0.266	92.7% (0.6%)	0.19
npcausal	-0.002 (0.002)	0.005	0.260	93.9% (0.5%)	0.98
tmle	-0.002 (0.002)	0.005	0.257	94.0% (0.5%)	0.86
tmle3	-0.002 (0.002)	0.005	0.261	93.9% (0.5%)	4.54
GAMs + k=10 sample splitting					
AIPW	-0.002 (0.002)	0.005	0.310	96.6% (0.4%)	7.92
npcausal	-0.002 (0.002)	0.006	0.319	96.5% (0.4%)	3.55
tmle	-0.002 (0.002)	0.005	0.272	95.6% (0.5%)	5.15
tmle3	-0.002 (0.002)	0.005	0.308	96.5% (0.4%)	7.51
SuperLearner + No sample splitting					
AIPW	-0.009 (0.002)	0.005	0.246	93.0% (0.6%)	14.65
npcausal	-0.005 (0.002)	0.005	0.232	90.3% (0.7%)	21.71
tmle	-0.009 (0.002)	0.005	0.251	93.8% (0.5%)	13.44
tmle3	-0.005 (0.002)	0.005	0.246	92.2% (0.6%)	36.76
SuperLearner + k=10 sample splitting					
AIPW	-0.002 (0.002)	0.005	0.281	95.6% (0.5%)	128.48
npcausal	-0.004 (0.002)	0.005	0.285	95.5% (0.5%)	183.54
tmle	-0.006 (0.002)	0.005	0.266	94.5% (0.5%)	43.38
tmle3	-0.004 (0.002)	0.005	0.272	95.2% (0.5%)	48.52

* Cross-fitting was conducted in the outcome model only because of its implementation.

[1] Sample size (n) = 200; Number of simulation (nSim) = 2000; RD_{true} = 0.128; Numbers within parentheses are Monte Carlo SEs of the performance indicator estimates

[2] Asymptotic SEs were used for CI calculation in AIPW, CausalGAM, tmle and tmle3. CIs for gComp and IPW were obtained by 200 bootstraps and sandwich estimators, respectively

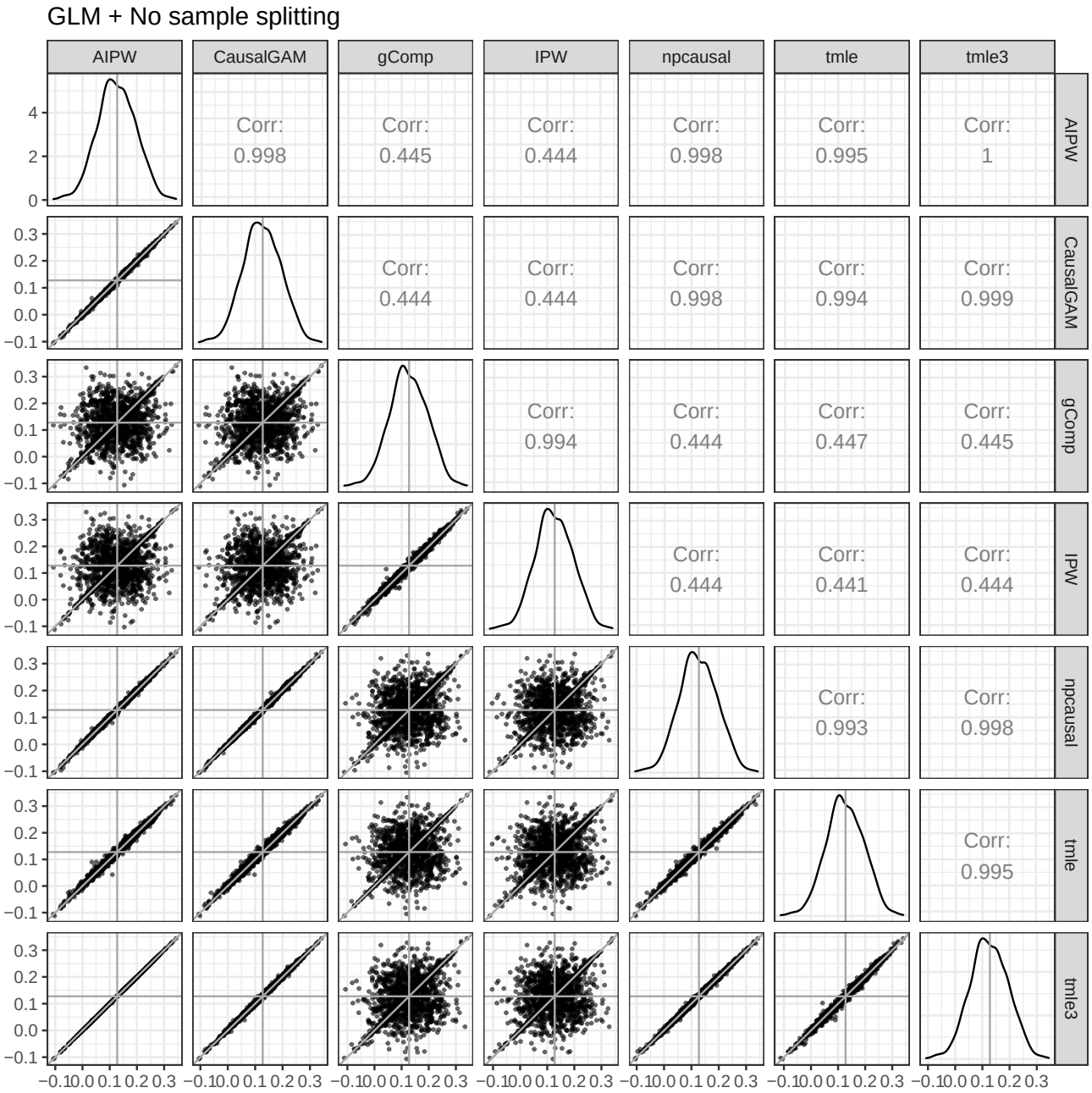
[3] SuperLearner was used for tmle and AIPW and sl3 for tmle3; Algorithms include GAMs, earth, ranger and xgboost

Table 7: Comparisons of R packages implementing doubly robust (DR) estimators

Packages	AIPW	CausalGAM	npcausal	tmle	tmle3
Version evaluated	0.6.3.1	0.1-4	0.1.0	1.4.0.1	0.1.7
DR estimator	AIPW	AIPW	AIPW	TMLE	TMLE
Available model	Super Learner	GAMs	Super Learner	Super Learner	Super Learner
Cross-fitting	Yes	No	Yes	Yes	Yes
Different covariate sets	Yes	Yes	Yes[1]	Yes[2]	Yes
Exposure type	Binary[3]	Binary	Binary, Categorical, Continuous	Binary[3]	Binary, Categorical, Continuous
Propensity score truncation	Yes	Yes	No	Yes	Yes
Outcome type	Binary & Continuous	Binary & Continuous	Binary & Continuous	Binary & Continuous	Binary & Continuous
Missing data support	Missing outcome	No	No	Missing outcome	Missing outcome
ATE estimate scale	RD, RR, OR	RD	RD	RD, RR, OR	RD, RR, OR
SE type	Asymptotic	Asymptotic, Sandwich, Bootstrap	Asymptotic	Asymptotic	Asymptotic
Parallel processing	Yes	No	No	No	Yes

[1] Users need to manually input propensity scores for different covariate sets.
[2] When different covariate sets is enabled, tmle only supports glm for estimation.
[3] Continuous and categorical exposures can be used but need to be binarized.[75]

Figure 7: Pairwise comparison of ATE estimates with the true data generating functions using different methods



Diagonal panels are the density plots of estimates from each package, lower diagonal panels are scatter plots of estimates between two packages, and upper diagonal panels are Pearson correlations of estimates between two packages. In the scatter plots, horizontal and vertical lines refer to RD_{true} = 0.128, and diagonal lines are references with a slope = 1 and an intercept of 0

4.0 Estimating Per-Protocol Treatment Effects Using Machine Learning in Randomized Clinical Trials

4.1 Introduction

Intention-to-treat (ITT) effects from randomized controlled trials (RCT) are the gold standard for evaluating treatment effects. Importantly, ITT effects capture the impact of assigning treatments to individuals. The ITT approach does not provide estimates of the effects that would be observed if all individuals adhered with a desired treatment protocol.[16, 17, 25] That is, in the presence of non-adherence, the ITT effects may differ in important ways from the effect of taking the treatment under study in a specified way (protocol).[17, 15]

Recently, several researchers have called for formal per-protocol analyses of randomized trials,[17, 138] and several recent per-protocol analyses have demonstrated important deviations from the ITT estimates when non-adherence is accounted for.[15, 139, 140, 141, 142, 143, 144, 145] Unfortunately, when per-protocol effects are targeted in randomized trials, all of the limitations associated with observational studies must be considered, such as confounding bias.[17, 15] Machine learning methods can be used with contemporary causal inference methods to overcome some of these limitations,[49, 50] and estimate per-protocol effects adjusting for confounding variables. However, compared to traditional regression models, machine learning methods may be better suited to avoiding problems with model mis-specification.[146, 121] For example, a model would be mis-specified if a linear regression were used to fit two variables with non-linear relations (e.g., perinatal mortality and maternal hemoglobin).[147] Many machine learning algo-

rithms can avoid these problems,[146, 121] but they have not yet been applied to scenarios where per-protocol effects are of primary interest.

In this paper, we illustrate the use of machine learning methods to estimate the per-protocol effects of low-dose aspirin on pregnancy in the Effects of Aspirin in Gestation and Reproduction trial. We demonstrate how machine learning methods can be used to estimate per-protocol effects with causal inference methods, and discuss the feasibility and trade-offs of using machine learning methods for adherence-adjusted analyses.

4.2 Method

4.2.1 Study Design

We used the data from the Effects of Aspirin in Gestation and Reproduction (EAGeR) trial (ClinicalTrials.gov: NCT00467363), a multicenter, block randomized, double-blind, placebo-controlled trial. EAGeR recruited women aged 18-40 years who were actively trying to become pregnant with one or two prior pregnancy loss and no history of infertility from four university medical centers in the United States from 2007 to 2012. A total of 1228 were recruited and randomized. For up to six menstrual cycles, participants were followed biweekly in their first two cycles and monthly afterwards while attempting pregnancy. If a pregnancy was observed, follow-up continued throughout pregnancy for the live birth outcome (the registered primary endpoint of the trial). Institutional review board approvals at each clinical site and data coordinating center were obtained. A Data Safety and Monitoring Board was also formed to ensure participants' safety and monitor the efficacy of the trial. Missing data were addressed via single imputation. Details about

study design, eligibility criteria, baseline characteristics and other relevant information has been published elsewhere.[12, 13, 14, 15]

4.2.2 Treatment and Adherence

With 1:1 randomization allocation, the treatment group (n=615) received preconception-initiated daily low-dose aspirin (81 mg) plus folic acid (400 µg) and the control group (n=613) received placebo plus folic acid (400 µg). For women who became pregnant, study treatment was to continue until the 36th week of gestation. A total of 1227 women were included in the analysis of the current study (one participant has missing follow-up data).

Adherence was assessed via bottle weight measurements in both groups during regular follow-up visits. Weekly adherence status was determined by evaluating whether a participant took their assigned pills for at least 5 out of 7 days (equivalent to 70%) during a given week. A woman was deemed adherent with our study protocol if she took 5/7 pills for at least 80% of their follow-up time before becoming pregnant, or entire follow-up time for those without pregnancy. Notably, this adherence status is a dichotomized, time-fixed variable, which is commonly used in a typical per-protocol analysis but differs from the previous analyses of this trial.[15]

4.2.3 Outcome

Human chorionic gonadotropin (hCG) detected pregnancy during the defined treatment period was the primary outcome for this analysis. Pregnancies were determined by a positive result on a "real-time" hCG pregnancy test (Quidel Quickvue), which was sensitive to 25 mIU/ML hCG. The test was conducted at each study visit when expected

menses are absent, or by batched urine hCG testing using daily first-morning urine collected at home, stored on the last 10 days of each participant's first cycle after randomization, and analyzed in the laboratory.

4.2.4 Baseline Covariates and Post-randomization Confounders

Baseline data on demographic, behavioral and pregnancy history information was obtained by questionnaires, including age, race, education, marital status, income, exercise, alcohol and cigarette use in the past year, number of prior pregnancy losses and number of months attempting pregnancy prior to randomization. Physical measurements of height and weight were used to calculate body mass index (BMI) at baseline. Blood samples were also collected to measure serum high-sensitivity C reactive protein (hsCRP), using an immunoturbidmetric assay (Roche COBAS 6000 autoanalyzer) with a detection limit of 0.015 mg/L.

Post-randomization confounders, including unusual (or excessive) bleeding and nausea and/or vomiting were collected via questionnaire at regular intervals over the course of follow-up. Similar to overall adherence status, we dichotomized these two post-randomization confounders by setting the values to 1 if a woman experienced unusual bleeding, or nausea and/or vomiting $\geq 1/7$ days (20%) per week for at least 50% and 20% of their follow-up time, respectively.

4.2.5 Statistical Analysis

In this study, we selected a protocol in which women would adhere to their assigned treatment for at least 5/7 days of a given week and for over 80% of the time in which they were followed before pregnancy. This allows us to evaluate whether consistently taking

aspirin (versus placebo) has any beneficial impact on the probability of experiencing an hCG-detected pregnancy. To examine the impact of different adherence thresholds on the overall findings, as well as the number of adherent and non-adherent individuals in the samples, we explored protocols under assigned treatment for at least 4/7, 5/7, 6/7 days of a given week over 60%, 70%, 80% person-time.

To estimate the per-protocol effect of interest with machine learning methods, we used an augmented inverse probability weighting (AIPW) estimator with an ensemble machine learner known as the Super Learner (or stacked generalization).[51, 122, 121, 77] Per-protocol effects were quantified on both the risk difference (RD) and the risk ratio (RR) scales for the dichotomous pregnancy outcome.

Stacking is a machine learning technique that combines several different algorithms into a single "meta-algorithm". The benefit of using stacking, as opposed to a single regression model or machine learning algorithm (e.g., the LASSO regression or random forests), is flexibility: stacking algorithms can combine the strengths of each individual algorithm based on how they fit the data, thus avoiding the need of the potentially strong assumptions that single algorithms rely on for validity. The stacking technique first trains several machine learning models individually as the "first" layer. Estimates (or predictions) of the individual models from the first layer are then used as the input for the "second" layer which is the meta-algorithm. Cross-validation is used to determine the importance of each first-layer algorithm in the overall meta-algorithm, and to avoid potential overfitting.[122, 121]

In this study, we stacked five regression models (from traditional to flexible): a standard generalized linear model with main effects only, a standard generalized linear model with main effects and all two-way interactions, multivariate adaptive regression splines

(MARS),[124] random forests,[148] and extreme gradient boosting.[126] For MARS, random forest, and extreme gradient boosting, a grid of tuning parameters was included in the stacking algorithm. All algorithms were combined into the meta-algorithm via non-negative least squares. The predictions from these stacked models were then used to construct the augmented IP weighting estimator.

Augmented IP weighting is a so-called "doubly robust" estimator that relies on estimating the exposure model (i.e., propensity score) and the outcome model separately (both modeled with the stacking algorithm), and then combining the predictions from these models into a single estimator that quantifies the average treatment effect.[51] Augmented IP weighting is consistent as long as at least one of the exposure model or the outcome model is correctly specified. Further, augmented IP weighting performs well (i.e., parametric 1/n mean squared error and closed-form confidence intervals), even when using flexible machine learning methods.[49, 77] Using the stacked machine learning algorithm describe above, we estimated propensity scores by modeling the exposure with the aforementioned baseline covariates (exposure model) and constructed the outcome model using the exposure and those covariates. Cross-fitting, an additional layer of fitting process on top of the stacking machine learning, is applied in the augmented IP weighting estimator to obtain valid inference (e.g., low bias) and to further avoid over-fitting.[49, 77]

Sensitivity analyses were conducted by using other thresholds of time-fixed adherence status, which is a combination of adhered to at least 4, 5, 6 days (60%, 70%, 80%) in a given week over at least 60%, 70%, 80% person-week of follow-up. In addition, we also provided the ITT estimate, unadjusted per-protocol effects (with different thresholds) as well as the per-protocol effects estimated by g-computation,[76] inverse probability (IP) weighting[74] and targeted maximum likelihood estimation (TMLE),[149] re-

spectively. G-computation and IP weighting were constructed with standard generalized linear model with main effects only. TMLE is also a doubly robust estimator, which performs well when machine learning methods are used. We constructed the TMLE estimator using the same stacking machine learning algorithms for the augmented IP weighting. Further, we repeated all analyses adjusting for post-randomization confounders (i.e., unusual bleeding, nausea and/or vomiting).

All analyses were performed in R (3.6.2). We conducted the augmented IP weighting estimation using the *AIPW* package. The *AIPW* package supports the *SuperLearner* package for stacking machine learning with cross-validation, and provides a user-friendly interface for cross-fitting.[130, 77] A prior simulation study using the data resampled from EAGeR has shown excellent statistical performance for the AIPW package.[77] TMLE was conducted with the *tmle* package.[75] The code needed to reproduce our analyses is available in the Appendix C.2.

4.3 Results

Figure 8 presents the number of participants who adhered to the protocol, which decreases as the adherence threshold increases. Overall, 858 (70.0%) of the 1227 trial participants were adherent to their assigned study medication, and 784 (63.9%) of participants became pregnant. Table 8 shows the randomized treatment assignment, outcome, baseline characteristics and post-randomization confounders by adherence status. Adhering to at least 5/7 pills in a given week over at least 80% person-time was statistically associated with the hCG-detected pregnancy outcome as well as non-Hispanic White race/ethnicity, high school education, marital status, annual income, history of

smoking in the past year, and hsCRP at baseline, but not with the randomized treatment assignment.

The estimated per-protocol effect of low-dose aspirin on hCG-detected pregnancy is shown in Table 9. Among those who adhered to the treatment protocol, low-dose aspirin increased the probability of hCG-detected pregnancy by 0.080 (95% Confidence Interval or CI, 0.025 to 0.136). This per-protocol risk difference of low-dose aspirin is about double the ITT estimate (0.043, 95%CI, -0.011 to 0.096). Similar per-protocol effects were also observed when adjusting for unusual bleeding, and nausea and/or vomiting (RD: 0.084, 95%CI, 0.028 to 0.140). Risk ratios for the estimated per protocol effects are also presented in Table 9.

Using other adherence thresholds, our sensitivity analyses with AIPW and machine learning show the per-protocol effects are larger as the adherence to assignment treatment increases. These estimated per protocol effects ranged from 0.056 (95% CI, 0.0001 to 0.112) to 0.090 (95%CI, 0.034 to 0.145) when adherence thresholds ranged from 4/7 days for at least 60% person-time of follow-up to 6/7 for at least 80% person-time (See Figure 20). Using other estimation methods, the per-protocol estimates remain similar, including the unadjusted estimates (Table 12 and Figure 20).

4.4 Discussion

We demonstrate the use of stacking machine learning with augmented IP weighting in estimating the per-protocol effects in the EAGeR trial. Our time-fixed per-protocol analyses results were consistent with previous findings of the per-protocol effect of aspirin that accounted for the time-varying nature of adherence and select time-varying

confounders.[15] However, unlike previous research, we used nonparametric machine learning methods to estimate these effects. Our analyses demonstrate a novel approach for per-protocol effect estimation using advanced statistical methods. In addition, our results suggest preconception low-dose aspirin increases hCG-detected pregnancies for women with one or two prior pregnancy losses, who adhered to 5/7 days of low-dose aspirin over 80% of the follow-up.

Supervised machine learning algorithms have been widely adopted to predict various health outcomes.[150, 151, 152] While they can also be used for effect estimation, additional steps are needed.[50, 49] Importantly, these include the need to adjust for relevant confounders, and to use doubly robust methods such as AIPW.

The benefits of using machine learning with doubly robust methods lie primarily in the ability to avoid strong parametric modeling assumptions. Machine learning models are more flexible and data-adaptive than traditional regression models for prediction.[146, 121] For example, whether to include an interaction term in a regression model is determined by the investigators' domain-specific knowledge, whereas tree-based models (e.g., random forests) adopt a more data-adaptive approach to interaction inclusion.[153, 154] Failure to include an interaction term may result in model mis-specification and lead to biased effect estimation. However, as a result of this increased data-adaptiveness and extra modeling flexibility, tree-based models, and flexible machine learning in general, are more likely to overfit the data and can suffer larger mean squared error.[153] To mitigate these issues, combining tree-based methods such as random forests, as well as regression based methods such as generalized linear regressions and MARS are advisable.[49, 121] In our study, we stacked five different machine learning models for an even more flexible algorithm and used cross-validation to mitigate overfitting.

We used supervised machine learning methods with doubly-robust estimators to quantify the per protocol effect of aspirin on hCG-detected pregnancy. In a randomized trial where all participants are fully adherent with the treatment protocol, per-protocol effects will be identical to ITT effects.[155] However, in the EAGeR trial, the ITT effects of low-dose aspirin on hCG-detected pregnancy differed substantially from the estimated per protocol effect due to non-adherence with the specified protocol over the course of follow-up. In many settings captured by pragmatic trials, perfect adherence is unlikely, and a practical adherence level has to be chosen based on either clinical knowledge or the data at hand. In EAGeR, the adherence rate declined overtime and dropped drastically after the start of pregnancy.[15] We chose a protocol of taking 5/7 pills in a given week over at least 80% person-time as the adherence level because existing literature suggests some biological effect of low-dose aspirin could be achieved at this adherence level,[15] and because of the relatively short half-life of aspirin.[156]

In a well-conducted trial, an ITT approach provides unbiased estimates of the assigned treatments on defined outcomes. The ITT estimates capture the impact of the "treatment strategy" and generally can be interpreted as the effectiveness of recommending or prescribing one treatment as compared to another. In contrast, an appropriately adjusted per protocol analysis can be used to estimate the effect of taking the active treatment according to the specifications allowing estimation of the treatment efficacy. Similar to most per-protocol analyses, our study relied on time-fixed adherence status, which is an important limitation. Although our effect estimates of low-dose aspirin on hCG-detected pregnancy are similar to the prior study that accounted for time-varying adherence,[15] limitations should be considered when conducting time-fixed per-protocol analysis. First, in conducting a time-fixed analysis, we had to collapse time-varying ad-

herence status into a single time-point, losing detailed information of how adherence changed over follow-up. Second, time-fixed analyses are generally unable to appropriately adjust for time-varying confounders, such as unusual bleeding and nausea. For example, at a given time-point, adherence to treatment is associated with an increased likelihood of side effects (e.g., unusual bleeding), which is further associated with a decreased of adherence at the next time-point. As such, post-randomization confounders such as unusual bleeding and nausea could simultaneously mediate and confound the effect of adherence status, requiring an analytic approach that we did not use.[76] In addition, other common limitations of observational studies should also be considered in the per-protocol analysis, such as unmeasured confounders.

Despite these limitations, we found that our unadjusted estimates are similar to estimates we obtained by improperly adjusting for post-randomization confounders (unusual bleeding, and nausea and/or vomiting), but properly adjusting for baseline confounders (e.g., age, marital status, annual income). Additionally, these results are closely aligned with results from a prior study where post-randomization confounding was properly adjusted for, albeit with methods that were much less flexible (parametric g computation).[15] This lends additional empirical support to the use of daily low-dose aspirin in increasing hCG detected pregnancies.

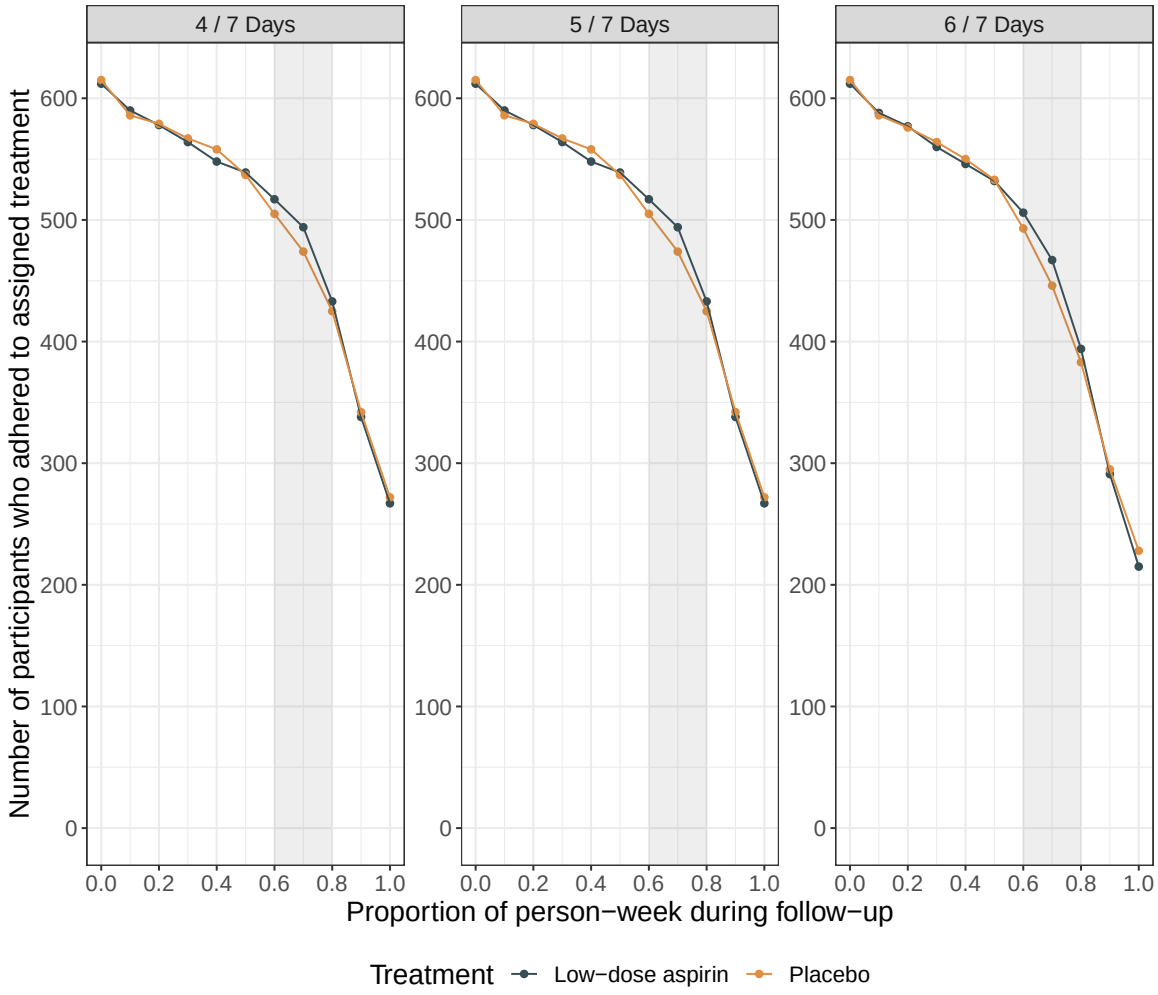
Our analytic approach using time-fixed adherence has a broad application in other analysis principles of RCT as well as in observational studies, especially for those with only one time-point. For example, our approach can be directly applied to the as-treated analysis in RCTs, such as a trial for evaluating the efficacy of emergency contraception. Modified ITT (despite not being consistently defined)[157, 158] can be incorporated with our approach as well, because the modification of ITT may not be free of confounding

(e.g., only including participants with drug initiation for a non-blinded study). Further, adjusting for covariates with machine learning in (modified) ITT analysis can improve statistical efficiency for higher precision of treatment effect estimates.[58]

In conclusion, machine learning methods with doubly robust estimators, such as AIPW, can be used to estimate per-protocol treatment effects. Assumptions are embedded implicitly and explicitly in different analytic plans. Health scientists should evaluate the benefits and disadvantages before implementing these advanced methods.

4.5 Tables and Figures

Figure 8: Number of participants adhered to assigned treatment by different thresholds



Participants at the shaded adherence levels [i.e., at least 4, 5, 6 days (60%, 70%, 80%) in a given week over at least 60%, 70%, 80% person-week of follow-up] were selected for the main per-protocol analysis and pertinent sensitivity analyses.

Table 8: Treatment assignment, outcome, baseline covariates and post-randomization confounders by adhering to 5/7 pills (70%) per week over 80% person-week of follow-up

Variables	Non-adherent N=369 (30.0%)	Adherent N=858 (70.0%)	p
Treatment			
Daily low-dose aspirin (%)	190 (51.5)	425 (49.5)	0.53
Outcome			
hCG-detected pregnancy (%)	107 (29.0)	677 (78.9)	<0.001
Baseline covariates			
Non-Hispanic White (%)	334 (90.5)	827 (96.4)	<0.001
High School Education (%)	302 (81.8)	755 (88.0)	0.004
Married (%)	312 (84.6)	811 (94.5)	<0.001
Employed (%)	283 (76.7)	636 (74.1)	0.34
Annual income (≥ 40000) (%)	213 (57.7)	608 (70.9)	<0.001
Exercise per week (%)			0.320
Low	106 (28.7)	216 (25.2)	
Moderate	140 (37.9)	360 (42.0)	
High	123 (33.3)	282 (32.9)	
Number of previous pregnancy loss (%)			0.61
1	125 (33.9)	278 (32.4)	
2	244 (66.1)	580 (67.6)	
Number of previous live birth (%)			0.11

0	173 (46.9)	352 (41.0)	
1	125 (33.9)	308 (35.9)	
2	68 (18.4)	179 (20.9)	
3	3 (0.8)	19 (2.2)	
Alcohol (ever consumed in past year) (%)	137 (37.1)	271 (31.6)	0.06
Tobacco (ever smoked past year) (%)	71 (19.2)	81 (9.4)	<0.001
Age (mean (SD))	2 28.41 (5.01)	28.88 (4.70)	0.12
Number of months attempting pregnancy prior to randomization (mean (SD))	4.25 (3.50)	3.93 (3.42)	0.14
BMI (mean (SD))	26.73 (6.52)	26.13 (6.57)	0.14
hsCRP (mean (SD))	3.34 (6.99)	2.58 (4.12)	0.02
Post-randomization confounders			
Unusual bleeding (%)	346 (93.8)	783 (91.3)	0.14
Nausea and/or vomiting (%)	69 (18.7)	138 (16.1)	0.26

Table 9: Effects of low-dose aspirin on hCG-detected pregnancy among women adhered to the assigned treatment: 5/7 pills (70%) per week over at least 80% person-week of follow-up

	RD				RR			
Method	Est.	SE	LCL	UCL	Est.	SE	LCL	UCL
Intention-to-treat	0.043	0.027	-0.011	0.096	1.069	0.043	0.982	1.163
Per-protocol analysis adjusted for baseline covariates								
Machine Learning + AIPW	0.080	0.028	0.025	0.136	1.107	0.036	1.032	1.188
Per-protocol analysis adjusted for baseline covariates and post-randomization confounders*								
Machine Learning + AIPW	0.084	0.029	0.028	0.140	1.113	0.037	1.036	1.196

* Adjusted for bleeding [$\geq 1/7 days (20\%)$ per week over $\geq 50\%$ person-week] and nausea and/or vomiting [$\geq 1/7 days (20\%)$ per week over $\geq 20\%$ person-week]

5.0 Conclusion

The primary objective of this dissertation was to develop, evaluate, and apply advanced causal inference methods and tools to estimate treatment effects for improving pregnancy outcomes. The methods and tools we evaluated show both challenges and promises of using artificial intelligence and machine learning methods in real epidemiologic data. This chapter is meant to summarize the key findings of this dissertation in the context of the Effects of Aspirin in Gestation and Reproduction (EAGeR) study, and to discuss the strengths and limitations as well as the public health implications and future directions for the use of advanced artificial intelligence and machine learning methods in reproductive/perinatal epidemiology and other areas of epidemiology.

5.1 Summary of Findings

Aim 1. Evaluate the performance of Bayesian networks in selecting covariate adjustment sets for estimating the effects of low-dose aspirin on pregnancy outcomes.

Causal discovery methods (or structural learning of Bayesian networks) underperform in selecting sufficient confounders in our simulations. These approaches have shown good performance in other setting, such as discovering biological pathways using genomics data.[78, 84, 85, 86, 87, 88] However, these data tend to be characterized as simple causal (or graphical) structures, such as random graphs (the occurrence of edges follows a distribution function).[89] To mimick the complexity of epidemiologic data, our simulation study used data resampled from the EAGeR trial to generate challenging analytic sce-

narios characterized by M-/butterfly- bias. The evaluations of the max-min hill-climbing causal (MMHC) discovery algorithm generally show poor performance, suggesting that causal discovery algorithms require more development and refinement prior to using them in epidemiologic settings. Specifically, we found low accuracy in selecting correct confounder adjustment sets that appropriately block all backdoor paths from the exposure to the outcome. Our finding suggests causal discovery methods should not be used in lieu of domain-specific knowledge for generating causal diagrams to select confounder adjustment sets.

Aim 2. Develop an R package for augmented inverse probability weighting (AIPW) to estimate the effects of low-dose aspirin on pregnancy outcomes.

An increasing number of recent studies suggest doubly robust estimators with cross-fitting should be used when estimating causal effects with machine learning methods. However, existing programs that implement doubly robust estimators do not all support machine learning methods and cross-fitting, or provide estimates on multiplicative scales. The newly developed **AIPW** package is a state-of-the-art R program, implementing augmented inverse probability weighting, for doubly robust estimation of average causal effects. The **AIPW** package addresses the limitations of existing programs that implements doubly robust estimators. Those features include the supports of stacking machine learning algorithms via SuperLearner and sl3, cross-fitting for valid effect estimates, and providing estimates on both additive and multiplicative scales, and flexible covariate adjustment for randomized controlled trials. Our simulation shows that the **AIPW** package yielded comparable performance to existing R packages that implement doubly robust estimators (e.g., tmle). We also found that cross-fitting substantially improves the performance of doubly robust estimators fit with machine learning algorithms. As a result, our

AIPW package with stacking machine learning methods is useful to estimate adherence-adjusted effects of low-dose aspirin on pregnancy outcomes.

Aim 3. Determine the adherence adjusted effects of low-dose aspirin on pregnancy outcomes with Bayesian networks and AIPW.

Based on the results in the previous aims, we use the **AIPW** package with ensemble machine learning to estimate the time-fixed adherence adjusted effect of low-dose aspirin on hCG-detected pregnancy. Machine learning methods with the AIPW estimator are more flexible to estimate per-protocol treatment effects than traditional parametric regression, because these methods are not subject to strict parametric assumptions. As expected, low-dose aspirin increases the probability of hCG-detected pregnancy by 9.3% (95%CI 6%–18%). Using a different set of assumptions, our time-fixed per-protocol estimation using machine learning with doubly robust yielded similar results to the prior research using parametric g-computation with time-varying adherence (RD: 8.0% vs. 7.8%).[15]. Therefore, results of this aim support the previous findings that daily low-dose aspirin increases pregnancy among women with one or two pregnancy losses who adhered to the aspirin protocol.

5.2 Strengths and Limitations

This dissertation has a number of important strengths for the practice of epidemiology. First, our evaluated methods and developed tools are state-of-the-art in epidemiology, statistics and machine learning. The status quo of effect estimation in epidemiology relies on domain-specific knowledge from DAG identification to model specification. This dissertation uses data-adaptive approaches in machine learning and artificial intelli-

gence, and provides practical guidance for epidemiologists to adopt these approaches in epidemiologic studies, outlining the assumptions embedded in their use, as well as the pros and cons of these approaches. Second, the simulations conducted in this dissertation were constructed to reflect the real epidemiologic scenarios as closely as possible. By adopting a plasmode simulation approach that relied on data from the EAGeR trial, our simulated datasets were better able to preserve real relations among certain variables, as opposed to simply simulating artificial data.[159] In addition, the causal diagrams used to simulate our data in this dissertation were complex, but are often reasonable assumed to exist in real epidemiologic data.[36, 72] Third, the **AIPW** package we developed has a broad application in various type of epidemiologic data. Using the causal diagrams depicted in Figure 5, the use-case of the package can be extended to both observational studies and randomized trials. Further, the **AIPW** package provides a user-friendly interface for the use of machine learning and cross-fitting, allowing applied scientists to use these sophisticated methods in the most appropriate way possible. Finally, the per-protocol effect estimation in this dissertation is conducted with alternative assumptions (e.g., time-fixed adherence, free from parametric assumptions). Hence, the results from our analyses provide additional evidence of the benefits of low-dose aspirin on pregnancy outcomes.

However, several limitations should be considered when interpreting the findings in this dissertation. First, only a limited number of causal discovery algorithms are evaluated in a limited set of causal diagrams. Therefore, our evaluations may not be generalizable to the latest causal discovery algorithms, such as continuous optimization with neural networks,[160] or to the data with simple causal structures (e.g., genomics data).[86] Second, the design of this dissertation assumes that the adherence status does not change

over time. As such, the tools and approaches evaluated and developed do not consider time-varying exposure and confounder, as well as time-to-event outcomes. [161, 162] Collapsing the time-varying adherence into a time-fixed status loses detailed information in the EAGeR trial. Although there are trade-offs between time-fixed and time-varying per-protocol analyses, time-varying setting is more preferable in most epidemiologic studies because failure to adjust for post-randomization confounders properly may lead to biased effect estimation.[114, 163]

5.3 Public Health Implications

Pregnancy loss is a common but severe complication among pregnant women, which is associated with fertility.[14] As a cheap, safe and generic over-the-counter medication, aspirin is very promising to increase pregnancy rate and to prevent adverse pregnancy outcomes among women at higher risk of pregnancy loss. This dissertation provides additional evidence that low-dose aspirin can improve pregnancy rate among women who had one or two prior losses. Further, the current work developed the state-of-the-art tools and methods to determine the benefits of aspirin on pregnancy outcomes, yielding solutions to alleviate the public health burden of pregnancy loss.

Machine learning and artificial intelligence have geared epidemiologists with new perspectives to better understand population health problems. However, given the development of machine learning are more rapid than many disciplines of health sciences, the features of these new methods as well as their feasibility, strength and limitations in public health sciences are not well understood. This dissertation serves as a analytic framework for using machine learning to estimate treatment effects, and for evaluating

advanced machine learning methods in epidemiologic studies. As a result, the current work pushed forward the understanding of machine learning in public health research.

5.4 Future Research Directions

This dissertation opens up several venues for future research, in both methodological and practical domains. We highlight several important questions remain unanswered in this dissertation, which are needed to be addressed in future studies.

First, doubly robust estimators with machine learning are should be evaluated and further optimized in longitudinal setting. To our knowledge, LTMLE is the only tool to conduct doubly robust estimation for longitudinal data.[161] Current implementations of machine learning estimators for longitudinal data treat both the confounding variables, as well as the time component nonparametrically. However, in studies with long follow-up periods, treating the time component nonparametrically can easily lead to scenarios where there is not enough data to support nonparametric inference. Roughly, as the number of time-points increases, the number of variables also increases, leading to precipitous declines in precision. Hence, more work is needed to better understand the trade-offs between smoothing effect estimates across time, versus full nonparametric treatment of the time component.

Second, transportability, generalizability and data fusion of complex cohorts catch recent attentions of methodological research.[164, 165, 166, 167] For example, the findings from EAGeR may not be directly transported to another populations (e.g., women without prior pregnancy loss).[168, 169] However, current methodological studies on transportablity mainly focus on the use of parametric regressions.[166] As such, the applica-

bility of machine learning in these problems should be studied in the future.

Finally, high dimensional data provide opportunities in understanding the effectiveness of medications to improve pregnancy outcome.[93, 170, 171, 151, 172] For example, electronic health records have a tremendous amount of data, where machine learning plays a critical role in predicting health outcomes.[151, 172] The intersection of causal inference and machine learning is an active field of methodological studies. Therefore, the tools and methods in this dissertation need to be evaluated and adapted to high dimensional data for a boarder impact on population health research.

Appendix A Causal Discovery Appendix

A.1 Average treatment effect estimators

Under consistency, exchangeability, positivity, and no interference, the average of potential outcomes that would be observed under $A = a$ are identified as the average of estimated outcomes, that is: $E(Y^a) = E[E(Y \mid A = a, W)]$, which we denote as $\psi(a)$.

The inverse probability weighting (IPW) estimator can be constructed from propensity scores by modeling the exposure A as a function of confounders C:[74]

$$\hat{\psi}_{IPW}(a) = \left[\sum_{i=1}^{N} \frac{I(A_i = a)Y_i}{\hat{P}(A = a \mid C_i)}\right] \Bigg/ \left[\sum_{i=1}^{N} \frac{I(A_i = a)}{\hat{P}(A = a \mid C_i)}\right] \tag{7}$$

where $a \in \{0, 1\}$ and i represents i^{th} observation.

Counterfactual predictions $P(Y = 1 \mid A := a, C)$ can be used to construct a g-computation estimator:[76, 24]

$$\hat{\psi}_{gComp}(a) = \frac{1}{N} \sum_{i=1}^{N} \hat{P}(Y = 1 \mid A := a, C_i) \tag{8}$$

where the $:=$ symbol denotes that we set each individual's value for A in the sample to the argument's value a.

Further, both propensity scores and counterfactual predictions can be used to construct doubly robust estimators, such as augmented inverse probability weighting (AIPW):[114, 51, 77]

$$\hat{\psi}(a)_{AIPW} = \frac{1}{N} \sum_{i=1}^{N} \left[I(A_i = a)\frac{Y_i - \hat{P}(Y = 1|A_i, C_i)}{\hat{P}(A = a|C_i)} + \hat{P}(Y = 1|A := a, C_i)\right] \tag{9}$$

and targeted maximum likelihood estimation (TMLE) [75, 105] :

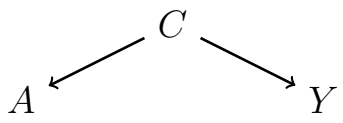
$$\hat{\psi}_{TMLE}(a) = \frac{1}{N} \sum_{i=1}^{N} \hat{P}^u(Y = 1 \mid A \coloneqq a, C_i) \tag{10}$$

where $\hat{P}^u$ is the updated probability of the counterfactual predictions using the "clever covariate", which is a function of the propensity score.

A.2 Introduction to the Max-Min Hill-Climbing (MMHC) causal discovery algorithm

Here, we describe one particular algorithm that is used for causal discovery, the Max-Min Hill-Climbing (MMHC) algorithm. Suppose we have a dataset with three discrete random variables $D : \{A, Y, C\}$ and a *large* number of observations, whose true causal structure can be represented with the following DAG G:

Figure 9: True causal DAG (G) for $D : \{A, Y, C\}$



Our objective is to use data D to recover the causal DAG G. In causal modeling, the causal DAG implies a factorization for the joint probability distribution of the variables in the DAG. Specifically, the joint probability distribution can be written as a product of the conditional distribution of each variable given its parents in the DAG. That is, each DAG can be represented by a corresponding structural equation model.[61] In the example in Figure 9, the true causal DAG (G) can be written as $P(A, Y, C) = P(C)P(A \mid C)P(C \mid Y)$. This also implies that A is independent of Y given C.

Causal discovery methods generally fall into two types of approaches for learning the causal DAG from data: (a) Constraint-based approaches try to identify a (family of) graph(s) that imply the conditional indpendencies that hold in the data, assessed by statistical tests of independence (e.g., the χ^2 test). (b) Score-based approaches attempt to identify graphs that "fit" the data well, using a Bayesian or a penalized likelihood criterion for structural equation models (or factorization) represented by the graph. The approaches

have pros and cons, with constraint-based methods being more scalable and better at recovering the DAG skeleton (edges without orientations), and score-based methods being better at orienting edges and less prone to statistical errors.[78]

MMHC combines the two approaches, and is therefore classified as a hybrid algorithm. The method has two phases: In the constraint-based phase, it uses uses a heuristic process to find out a set of variables that are (conditionally) associated with a target variable using hypothesis testing, and iterates until all variables have their own sets (so called candidate parents and children or CPC). As a result, this algorithm will provide the skeleton of the DAG (undirected graph) given each variable has a set of associated variables, right panel of Figure 10.

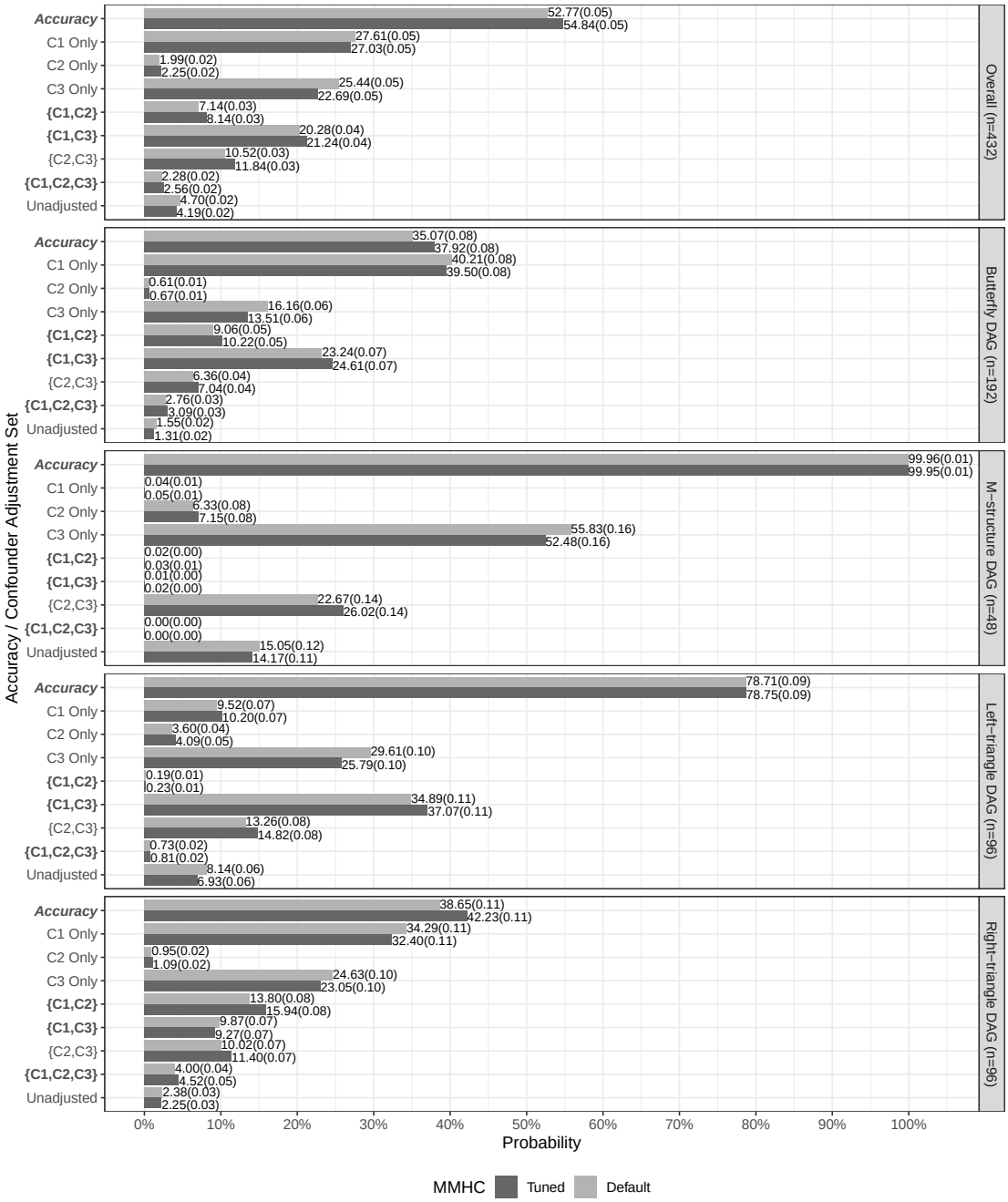
Figure 10: Simplified processes of MMHC algorithm for recovering G

In the score-based process, the MMHC algorithm starts with an empty graph to add, delete or reverse edges, and searches for the optimal scoring DAG. For example, Bayesian Information Criteria (BIC) can be calculated in the structural equation model of the DAG in each iteration. The algorithm uses a hill-climbing search to find a structural equation model that has an optimal BIC. To limit the search space and improve computational efficiency, the edge-adding process is restricted within the skeleton obtained from the constraint-based phase (i.e., the CPC). For example, the left-panel of Figure 10 represents the skeleton obtained from the constraint-based phase. The hill-climbing algorithm only can add edges between A, C and Y, C (i.e., no edge can be added between A, Y as shown

in the right-panel). The method will asymptotically discover the correct DAG, assuming no statistical errors.

A.3 Accuracy of MMHC algorithm in selecting correct confounder adjustment set

Figure 11: Accuracy of MMHC algorithm in selecting correct confounder adjustment set



- Numbers within parentheses are SE
- **Bold** sets are common admissible adjustment for all 432 DAGs.

Table 10: Accuracy of MMHC algorithm in selecting correct confounder adjustment set

			Default	Tuned
Num. DAGs	DAG type	Accuracy/ Adjustment Set	Prob (SE)	Prob (SE)
432	Overall	Accuracy	52.77% (0.05%)	54.84% (0.05%)
		Selected Adjustment Set		
		C1 Only	27.61% (0.05%)	27.03% (0.05%)
		C2 Only	1.99% (0.02%)	2.25% (0.02%)
		C3 Only	25.44% (0.05%)	22.69% (0.05%)
		{C1,C2}	7.14% (0.03%)	8.14% (0.03%)
		{C1,C3}	20.28% (0.04%)	21.24% (0.04%)
		{C2,C3}	10.52% (0.03%)	11.84% (0.03%)
		{C1,C2,C3}	2.28% (0.02%)	2.56% (0.02%)
		Empty/Unadjusted	4.70% (0.02%)	4.19% (0.02%)
192	Butterfly $\bowtie$	Accuracy	35.07% (0.08%)	37.92% (0.08%)
		Selected Adjustment Set		
		C1 Only	40.21% (0.08%)	39.50% (0.08%)
		C2 Only	0.61% (0.01%)	0.67% (0.01%)
		C3 Only	16.16% (0.06%)	13.51% (0.06%)
		{C1,C2}$^{\checkmark}$	9.06% (0.05%)	10.22% (0.05%)
		{C1,C3}$^{\checkmark}$	23.24% (0.07%)	24.61% (0.07%)
		{C2,C3}	6.36% (0.04%)	7.04% (0.04%)
		{C1,C2,C3}$^{\checkmark}$	2.76% (0.03%)	3.09% (0.03%)
		Empty/Unadjusted	1.55% (0.02%)	1.31% (0.02%)

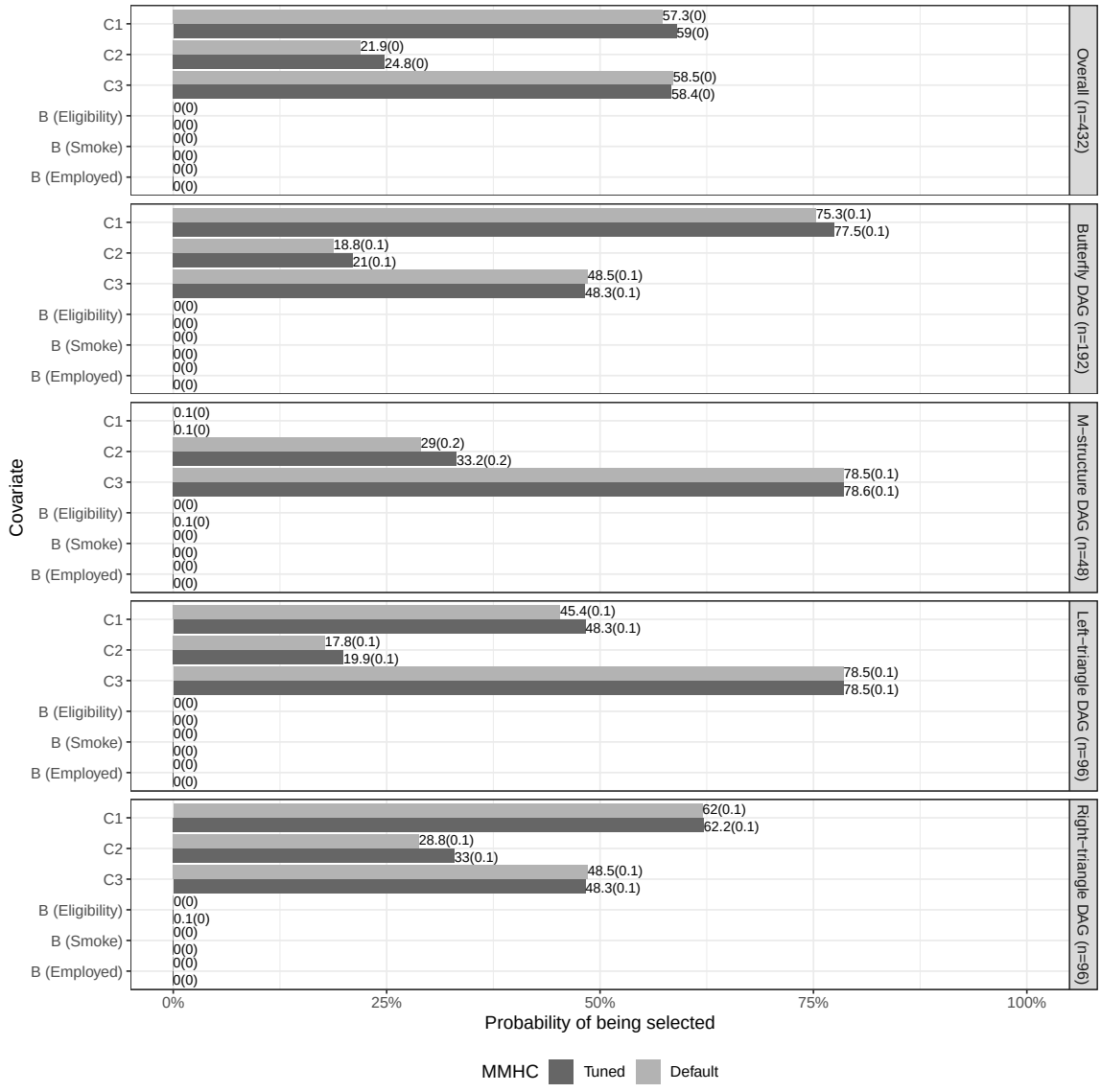
Num. DAGs	DAG type	Accuracy / Adjustment Set	Default Prob (SE)	Tuned Prob (SE)
48	M ⋈	Accuracy	99.96% (0.01%)	99.95% (0.01%)
		Selected Adjustment Set		
		C1 Only$^{\times}$	0.04% (0.01%)	0.05% (0.01%)
		C2 Only	6.33% (0.08%)	7.15% (0.08%)
		C3 Only	55.83% (0.16%)	52.48% (0.16%)
		{C1,C2}	0.02% (0.00%)	0.03% (0.01%)
		{C1,C3}	0.01% (0.00%)	0.02% (0.00%)
		{C2,C3}	22.67% (0.14%)	26.02% (0.14%)
		{C1,C2,C3}	0.00% (0.00%)	0.00% (0.00%)
		Empty/Unadjusted	15.05% (0.12%)	14.17% (0.11%)
96	Left-triangle ⋈	Accuracy	78.71% (0.09%)	78.75% (0.09%)
		Selected Adjustment Set		
		C1 Only	9.52% (0.07%)	10.2% (0.07%)
		C2 Only	3.60% (0.04%)	4.09% (0.05%)
		C3 Only$^{\checkmark}$	29.61% (0.10%)	25.79% (0.10%)
		{C1,C2}$^{\checkmark}$	0.19% (0.01%)	0.23% (0.01%)
		{C1,C3}$^{\checkmark}$	34.89% (0.11%)	37.07% (0.11%)
		{C2,C3}$^{\checkmark}$	13.26% (0.08%)	14.82% (0.08%)
		{C1,C2,C3}$^{\checkmark}$	0.73% (0.02%)	0.81% (0.02%)
		Empty/Unadjusted	8.14% (0.06%)	6.93% (0.06%)

Num. DAGs	DAG type	Accuracy / Adjustment Set	Default Prob (SE)	Tuned Prob (SE)
96	Right-triangle ⋈	Accuracy	38.65% (0.11%)	42.23% (0.11%)
		Selected Adjustment Set		
		C1 Only	34.29% (0.11%)	32.4% (0.11%)
		C2 Only✓	0.95% (0.02%)	1.09% (0.02%)
		C3 Only	24.63% (0.10%)	23.05% (0.10%)
		{C1,C2}✓	13.80% (0.08%)	15.94% (0.08%)
		{C1,C3}✓	9.87% (0.07%)	9.27% (0.07%)
		{C2,C3}✓	10.02% (0.07%)	11.40% (0.07%)
		{C1,C2,C3}✓	4.00% (0.04%)	4.52% (0.05%)
		Empty/Unadjusted	2.38% (0.03%)	2.25% (0.03%)

Probability and SE are calculated with n = number of DAGs * 2000 MC.
✓ Admissible adjustment set that blocks the backdoor path from A to Y
× Any adjustment set except the marked one.
Bold sets are common admissible adjustment for all 432 DAGs.

A.4 Probability of covariates selected by the default and the tuned MMHC, stratified by DAG type

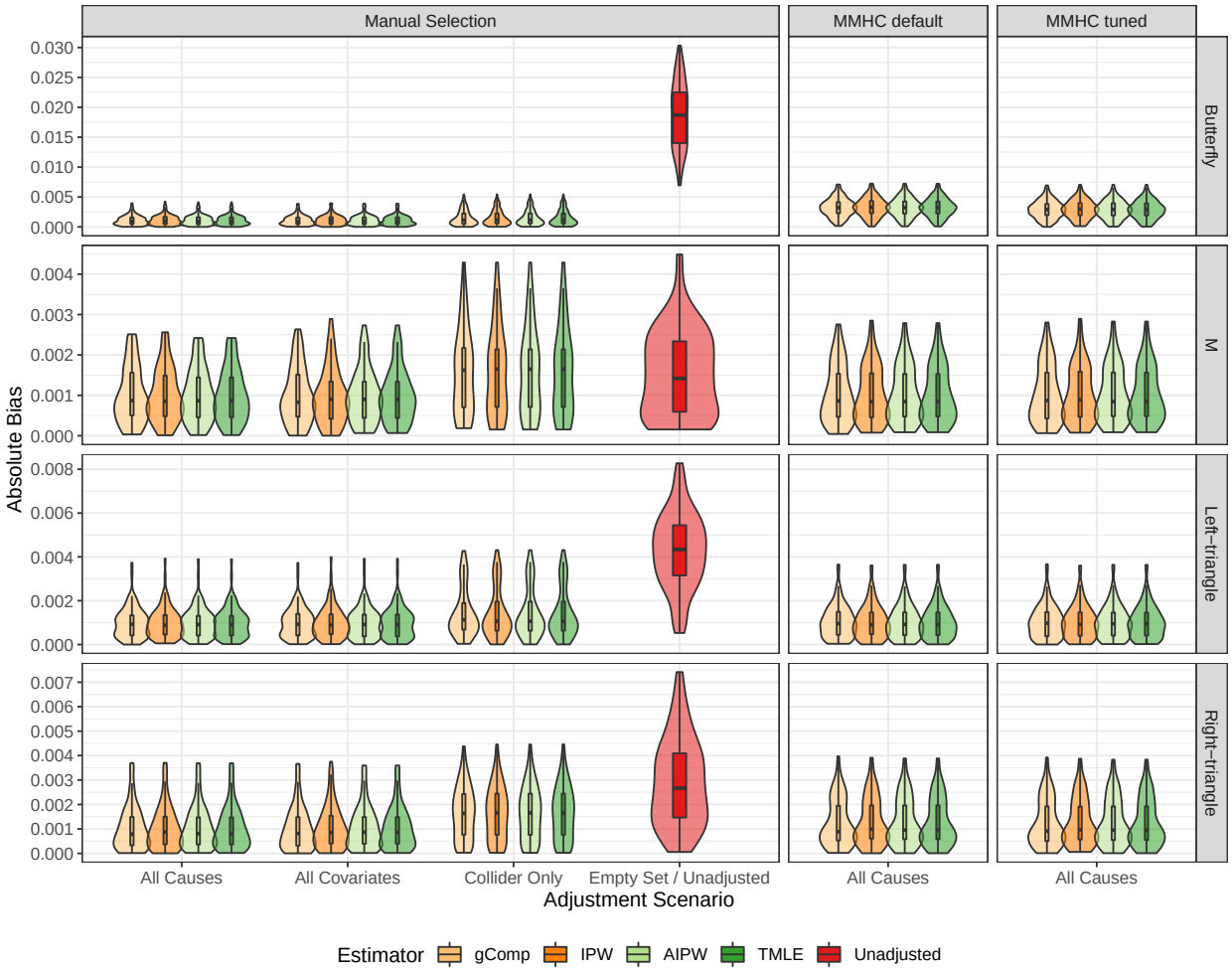
Figure 12: Probability of covariates selected by the default and the tuned MMHC, stratified by DAG type



- Numbers within parentheses are SE

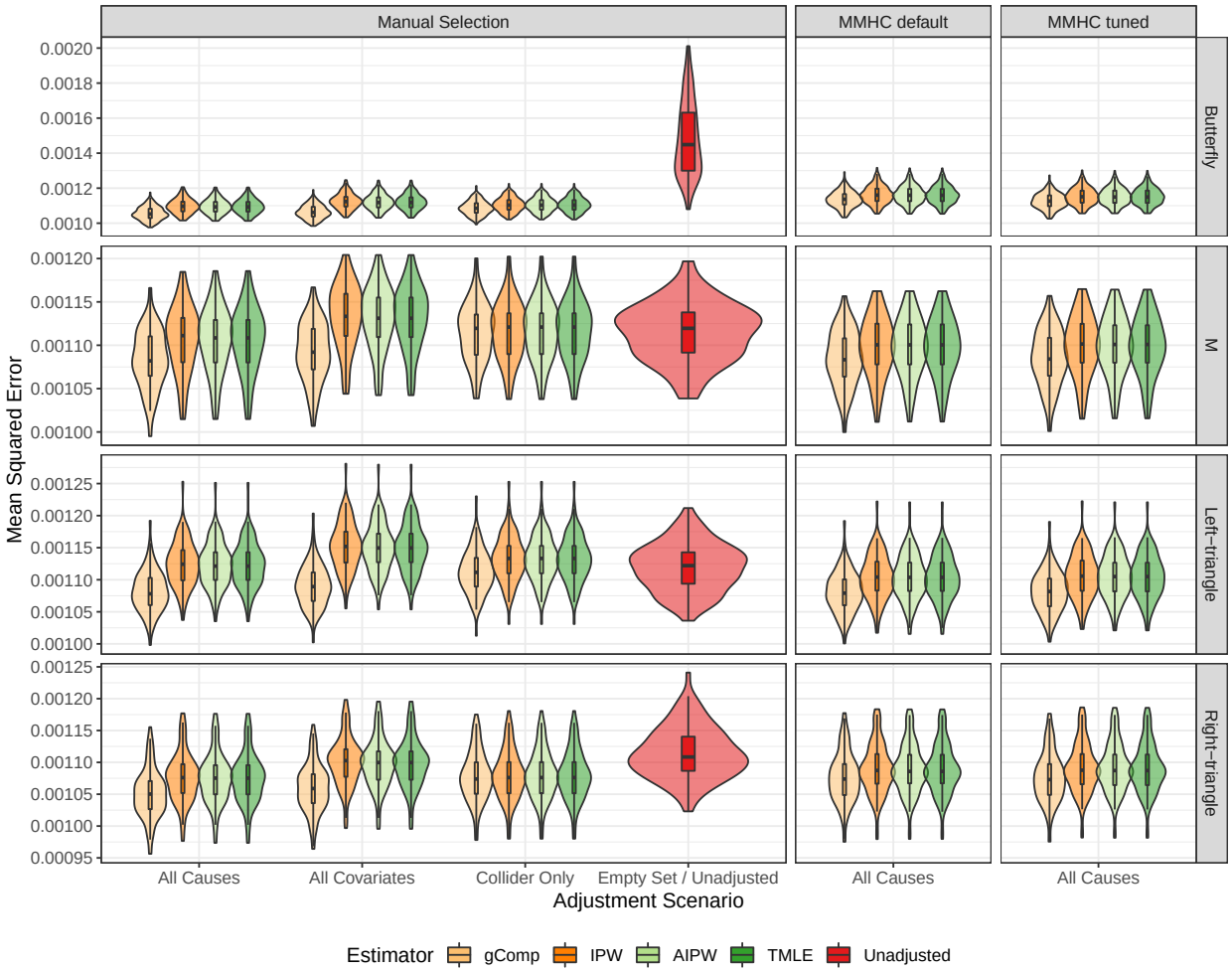
A.5 Distributions of absolute bias and mean squared error (MSE) of average treatment effect estimation

Figure 13: Distribution of absolute bias



- Each observation in the plot represents the absolute bias estimated in one of 432 DAGs using 2000 MC
- If MMHC algorithms yielded an empty confounder adjustment set, unadjusted estimates are used as the corresponding estimates from g computation, IPW, AIPW and TMLE.

Figure 14: Distribution of MSE



- Each observation in the plot represents the MSE estimated in one of 432 DAGs using 2000 MC
- If MMHC algorithms yielded an empty confounder adjustment set, unadjusted estimates are used as the corresponding estimates from g computation, IPW, AIPW and TMLE.

Appendix B AIPW Appendix

B.1 AIPW with missing outcome

Let R_i be an indicator of whether the outcome for individual i is observed ($R_i = 0$ if missing), and W be all of the covariates from W_Q and W_g. In the presence of missing outcome data, the AIPW estimator in the main text (formula 3) can be written as:

$$\hat{\psi}(a)_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{I(A_i = a, R_i = 1)}{\hat{P}(A = a, R = 1|W_i)}[Y_i - \hat{P}(Y = 1|A_i, W_i, R_i = 1)] + \hat{P}(Y = 1|A \coloneqq a, W_i, R_i = 1)\right\}$$

The propensity scores $\hat{P}(A = a, R = 1|W_i)$ is obtained by estimating the joint probability of treatment and (non)missingness:

$$\hat{P}(A = a, R = 1|W_i) = \hat{P}(R = 1|W_i, A = a)\hat{P}(A = a|W_i),$$

which incorporates missing data mechanism with W. In other words, analyses assume missing at random (MAR) conditional on W, and thus such analyses require W include covariates that render MAR as close to true as possible.

When missing outcomes are detected, the arguments in the AIPW package enabling different covariate sets for the outcome (W_Q) and exposure (W_g) models are disabled. This is because the propensity scores with (non)missing data can be factorized into two ways:

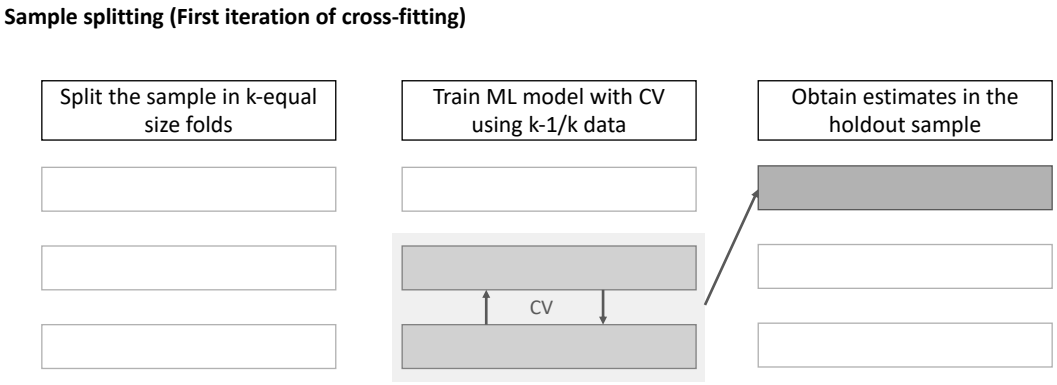
$$\begin{aligned}\hat{P}(A = a, R = 1|W_i) &= \hat{P}(R = 1|W_{Qi}, A = a)\hat{P}(A = a|W_{gi}) \\ &= \hat{P}(R = 1|W_{Qi})\hat{P}(A = a|W_{gi}, R = 1).\end{aligned}$$

In other words, it requires conditioning on both outcome covariates W_Q for missing data mechanism and W_g for exposure mechanism.

B.2 Implementation of sample splitting and cross-fitting

To implement sample splitting, one needs to subset the input data into k equal-size folds randomly, then fit the exposure and the outcome models with $(k-1)/k$ data, and finally use the fitted models to estimate propensity scores and outcome model predictions with the $1/k$ held-out sample.[127, 47] (Figure S1a)

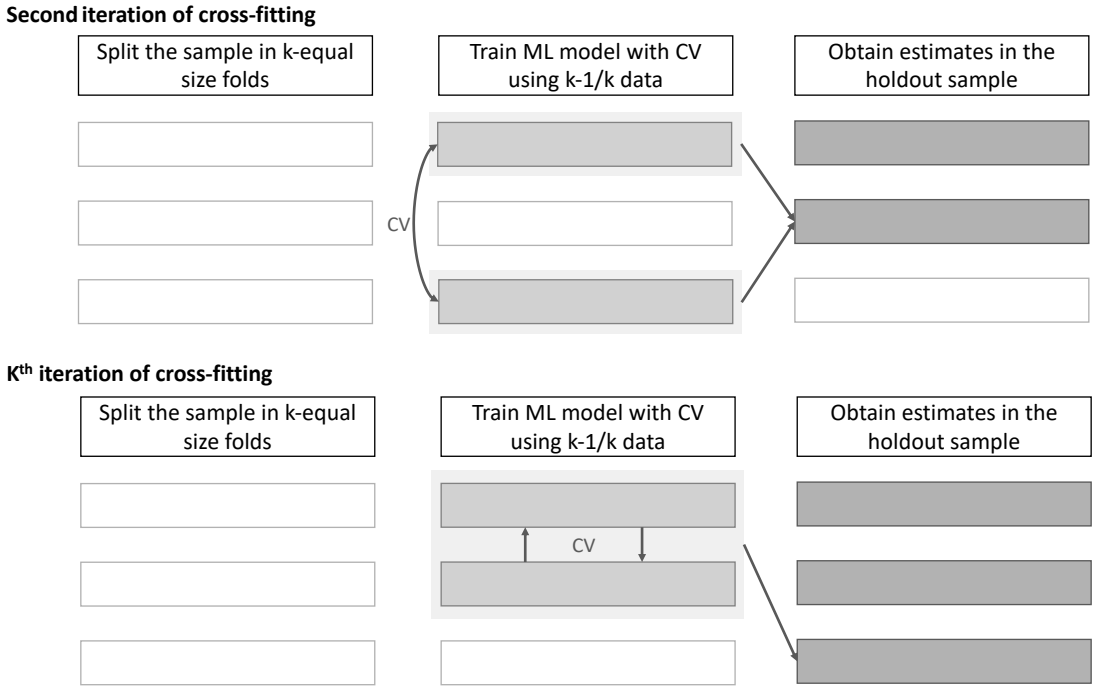
Figure 15: Illustration of sample splitting



ML: Stacking machine learning; CV: Cross-validation; k=3 in this example.

Cross-fitting is a more efficient version of sample splitting.[47] While sample splitting only uses $1/k$ of the sample for estimating propensity score and outcome model, cross-fitting iterates the process sample-splitting k times until estimates of the exposure and outcome for all observations are obtained.(Figure S1b).

Figure 16: Illustration of cross-fitting



ML: Stacking machine learning; CV: Cross-validation; k=3 in this example.

In the **AIPW** package with the **SuperLearner**, when $k\_split = 2$, 10-fold cross-validation (CV) will be used for training stacking machine learning algorithms; when $k\_split \geq 3$, $k\_split - 1$ fold CV will be used (e.g., 2-fold CV is used for $k\_split = 3$), with the CV-fold assignment remains the same throughout cross-fitting. With the **sl3** package, 10-fold CV will be used regardless of k_split.

B.3 Derivation of the standard errors of risk ratio and odds ratio for the AIPW estimator

Suppose we have an iid sample $Z_1, ..., Z_n \sim \mathbb{P}$ with $Z = (A, Y, W_Q, W_g)$ where $Y \in \{0, 1\}$. We assume the usual consistency, positivity, and no unmeasured confounding conditions.

Let

$$\widehat{\pi}(a \mid w_g) = \widehat{\mathbb{P}}(A = a \mid W_g = w_g) \text{ and } \widehat{\mu}(w_Q, a) = \widehat{\mathbb{P}}(Y = 1 \mid W_Q = w_Q, A = a)$$

denote estimators of the chance of receiving exposure level $A = a$ given covariate $W_g = w_g$, and the chance of observing outcome $Y = y$ among those with covariates $W_Q = w_Q$ and exposure $A = a$.

Under typical $n^{-1/4}$-type rate conditions, the following estimator is root-n consistent and asymptotically normal

$$\widehat{\mathbb{P}}(Y^a = 1) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\mathbb{1}(A = a)}{\widehat{\pi}(a \mid W_{gi})} \Big\{ Y_i - \widehat{\mu}(W_{Qi}, a) \Big\} + \widehat{\mu}(W_{Qi}, a) \right]$$

for the marginal counterfactual probability $\mathbb{P}(Y^a = 1) = \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}$. Note we use counterfactual expressions like $\mathbb{P}(Y^a = 1)$ as shorthand, but all the results here follow for the observational expressions $\mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}$ regardless of whether these equal the corresponding counterfactual expressions.

Therefore the following are estimators for the marginal risk ratio and odds ratio:

$$\widehat{\psi}_{rr} = \frac{\widehat{\mathbb{P}}(Y^1 = 1)}{\widehat{\mathbb{P}}(Y^0 = 1)}$$
$$\widehat{\psi}_{or} = \frac{\widehat{\mathbb{P}}(Y^1 = 1)/\{1 - \widehat{\mathbb{P}}(Y^1 = 1)\}}{\widehat{\mathbb{P}}(Y^0 = 1)/\{1 - \widehat{\mathbb{P}}(Y^0 = 1)\}}$$

Since both the RR and OR are non-negative, normal approximations will work best if we construct confidence intervals on the log scale and then exponentiate.

Let $\varphi_a(Z; \pi, \mu) = \frac{\mathbb{1}(A=a)}{\pi(a|W_{gi})}\Big\{Y_i - \mu(W_{Qi}, a)\Big\} + \mu(W_{Qi}, a)$ denote the uncentered influence function for $\mathbb{P}(Y^a = 1) = \mathbb{E}\{\mathbb{E}(Y \mid X, A = a)\}$ so that

$$\widehat{\mathbb{P}}(Y^a = 1) = \frac{1}{n}\sum_{i=1}^{n} \varphi_a(Z_i; \widehat{\pi}, \widehat{\mu})$$

Also let

$$\Sigma = \text{cov}\begin{pmatrix} \varphi_0(Z; \pi, \mu) \\ \varphi_1(Z; \pi, \mu) \end{pmatrix} = \begin{pmatrix} \text{var}\{\varphi_0(Z; \pi, \mu)\} & \text{cov}\{\varphi_0(Z; \pi, \mu), \varphi_1(Z; \pi, \mu)\} \\ \text{cov}\{\varphi_0(Z; \pi, \mu), \varphi_1(Z; \pi, \mu)\} & \text{var}\{\varphi_1(Z; \pi, \mu)\} \end{pmatrix}$$

denote the covariance matrix of the influence functions, with elements $\Sigma = \begin{pmatrix} \Sigma_{00} & \Sigma_{01} \\ \Sigma_{01} & \Sigma_{11} \end{pmatrix}$.

An estimate of the covariance matrix is simply given by

$$\widehat{\Sigma} = \begin{pmatrix} \widehat{\text{var}}\{\varphi_0(Z; \widehat{\pi}, \widehat{\mu})\} & \widehat{\text{cov}}\{\varphi_0(Z; \widehat{\pi}, \widehat{\mu}), \varphi_1(Z; \widehat{\pi}, \widehat{\mu})\} \\ \widehat{\text{cov}}\{\varphi_0(Z; \widehat{\pi}, \widehat{\mu}), \varphi_1(Z; \widehat{\pi}, \widehat{\mu})\} & \widehat{\text{var}}\{\varphi_1(Z; \widehat{\pi}, \widehat{\mu})\} \end{pmatrix}$$

where $\widehat{\text{cov}}$ and $\widehat{\text{var}}$ are just empirical covariances/variances.

Then under usual $n^{-1/4}$-type rate conditions on $(\widehat{\pi}, \widehat{\mu})$ we have

$$\sqrt{n}\left\{\begin{pmatrix}\widehat{\mathbb{P}}(Y^0=1)\\ \widehat{\mathbb{P}}(Y^1=1)\end{pmatrix} - \begin{pmatrix}\mathbb{P}(Y^0=1)\\ \mathbb{P}(Y^1=1)\end{pmatrix}\right\} \rightsquigarrow N(0,\Sigma)$$

Therefore by the delta method, we have

$$\sqrt{n}\Big(\log\widehat{\psi}_{rr} - \log\psi_{rr}\Big) \rightsquigarrow N\left(0, \begin{pmatrix}\frac{-1}{\mathbb{P}(Y^0=1)}\\ \frac{1}{\mathbb{P}(Y^1=1)}\end{pmatrix}^{\mathrm{T}} \Sigma \begin{pmatrix}\frac{-1}{\mathbb{P}(Y^0=1)}\\ \frac{1}{\mathbb{P}(Y^1=1)}\end{pmatrix}\right)$$

so that a 95% CI for ψ_{rr} is given by

$$\exp\left\{\log\widehat{\psi}_{rr} \pm \frac{1.96}{\sqrt{n}}\sqrt{\sum_{a=0}^{1}\frac{\widehat{\Sigma}_{aa}}{\widehat{\mathbb{P}}(Y^a=1)^2} - \frac{2\widehat{\Sigma}_{01}}{\widehat{\mathbb{P}}(Y^0=1)\widehat{\mathbb{P}}(Y^1=1)}}\right\}$$

Similarly the delta method also gives

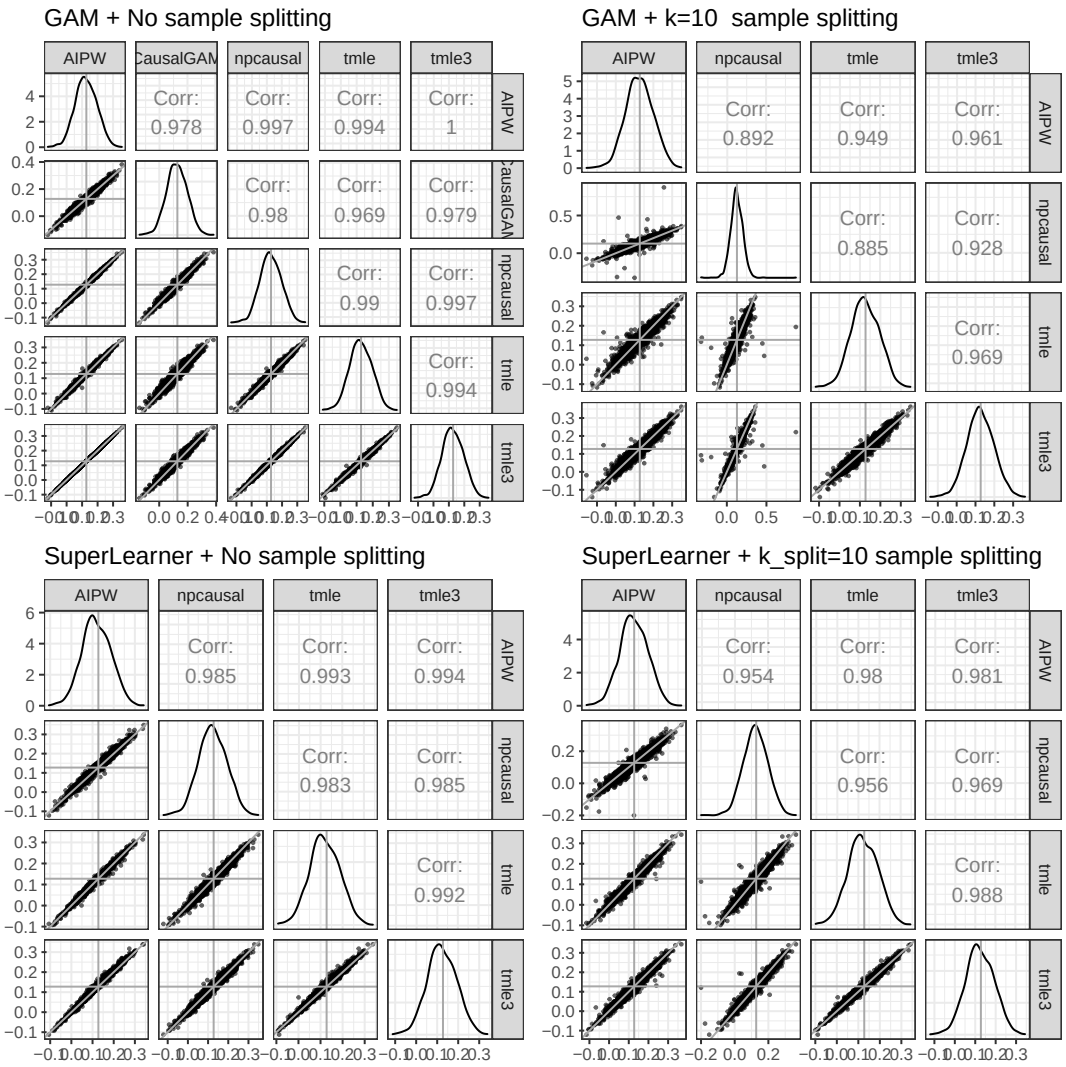
$$\sqrt{n}\Big(\log\widehat{\psi}_{or} - \log\psi_{or}\Big) \rightsquigarrow N\left(0, \begin{pmatrix}\frac{-1}{\mathbb{P}(Y^0=1)\mathbb{P}(Y^0=0)}\\ \frac{1}{\mathbb{P}(Y^1=1)\mathbb{P}(Y^1=0)}\end{pmatrix}^{\mathrm{T}} \Sigma \begin{pmatrix}\frac{-1}{\mathbb{P}(Y^0=1)\mathbb{P}(Y^0=0)}\\ \frac{1}{\mathbb{P}(Y^1=1)\mathbb{P}(Y^1=0)}\end{pmatrix}\right)$$

so that a 95% CI for ψ_{or} is given by

$$\exp\left\{\log\widehat{\psi}_{or} \pm \frac{1.96}{\sqrt{n}}\sqrt{\sum_{a=0}^{1}\frac{\widehat{\Sigma}_{aa}}{\widehat{\mathbb{P}}(Y^a=1)^2\widehat{\mathbb{P}}(Y^a=0)^2} - \frac{2\widehat{\Sigma}_{01}}{\widehat{\mathbb{P}}(Y^0=1)\widehat{\mathbb{P}}(Y^0=0)\widehat{\mathbb{P}}(Y^1=1)\widehat{\mathbb{P}}(Y^1=0)}}\right\}$$

B.4 Pairwise comparison of RD estimates using doubly robust packages with different estimation methods

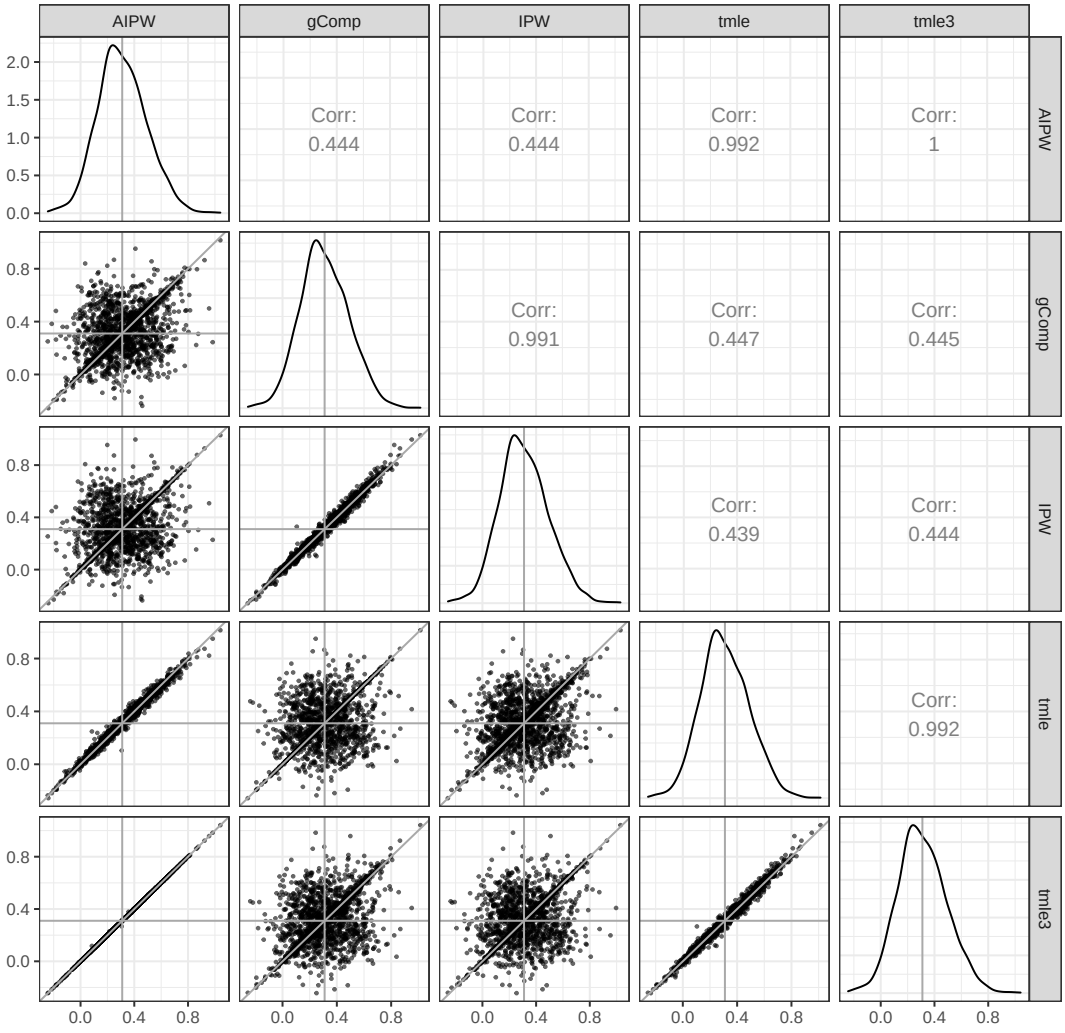
Figure 17: Pairwise comparison of RD estimates using doubly robust packages with different estimation methods



Diagonal panels are the density plots of estimates from each package, lower diagonal panels are scatter plots of estimates between two packages, and upper diagonal panels are Pearson correlations of estimates between two packages. In the scatter plots, horizontal and vertical lines refer to RD_{true} = 0.128, and diagonal lines are references with a slope = 1 and an intercept of 0

B.5 Pairwise comparison of log(RR) estimates with the true data generating functions using different methods

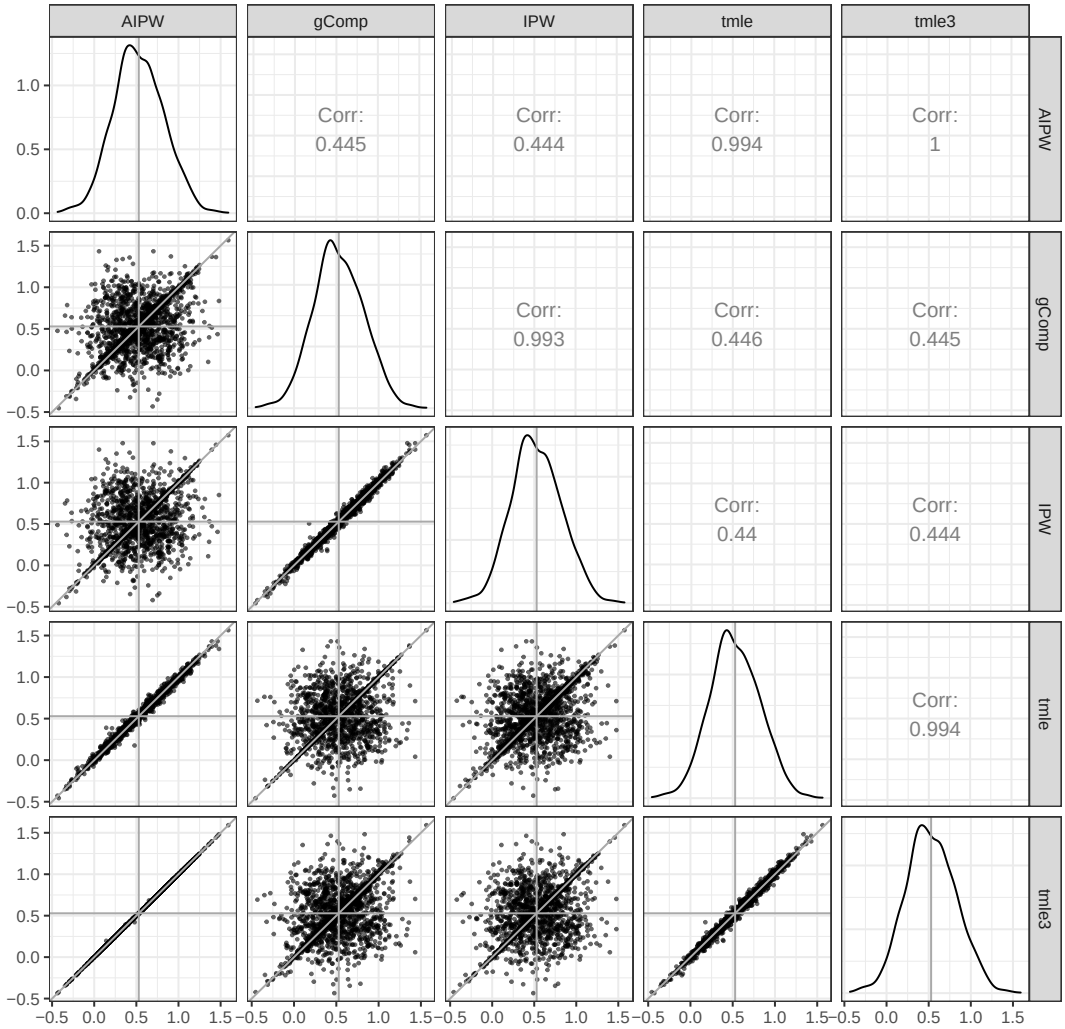
Figure 18: Pairwise comparison of log(RR) estimates with the true data generating functions using different methods



Diagonal panels are the density plots of estimates from each package, lower diagonal panels are scatter plots of estimates between two packages, and upper diagonal panels are Pearson correlations of estimates between two packages. In the scatter plots, horizontal and vertical lines refer to $log(RR_{true})$ = 0.31, and diagonal lines are references with a slope = 1 and an intercept of 0

B.6 Pairwise comparison of log(OR) estimates with the true data generating functions using different methods

Figure 19: Pairwise comparison of log(OR) estimates with the true data generating functions using different methods



Diagonal panels are the density plots of estimates from each package, lower diagonal panels are scatter plots of estimates between two packages, and upper diagonal panels are Pearson correlations of estimates between two packages. In the scatter plots, horizontal and vertical lines refer to $log(OR_{true})$ = 0.53, and diagonal lines are references with a slope = 1 and an intercept of 0

B.7 Performance of the AIPW package in estimating the average treatment effect [log(RR) and log(OR)]

Table 11: Performance of the AIPW package in estimating the average treatment effect [log(RR) and log(OR)] using true GLM model without cross-fitting in a simulated observational study based on EAGeR

Package/Method	Bias (SE)	MSE	MeanCIwidth	Coverage (SE)
log(RR)				
gComp	0.001 (0.004)	0.032	0.707	96.2% (0.4%)
IPW	0.001 (0.004)	0.033	0.719	96.4% (0.4%)
AIPW	0.001 (0.004)	0.033	0.675	94.8% (0.5%)
tmle	0.001 (0.004)	0.032	0.671	94.8% (0.5%)
tmle3	0.001 (0.004)	0.033	0.687	95.0% (0.5%)
log(RR)				
gComp	-0.002 (0.007)	0.087	1.171	95.7% (0.5%)
IPW	-0.002 (0.007)	0.090	1.195	96.1% (0.4%)
AIPW	-0.002 (0.007)	0.090	1.119	94.8% (0.5%)
tmle	-0.002 (0.007)	0.087	1.114	94.8% (0.5%)
tmle3	-0.002 (0.007)	0.090	1.141	95.1% (0.5%)

[1] Sample size (n) = 200; Number of simulation (nSim) = 2000; $log(RR_{true})$ = 0.31; $log(OR_{true})$ = 0.53; Numbers within parentheses are Monte Carlo SEs of the performance indicator estimates

[2] Asymptotic SEs were used for CI calculation in AIPW, tmle and tmle3. CIs for gComp and IPW were obtained by 200 bootstraps and sandwich estimators, respectively

Appendix C Effect Estimation Appendix

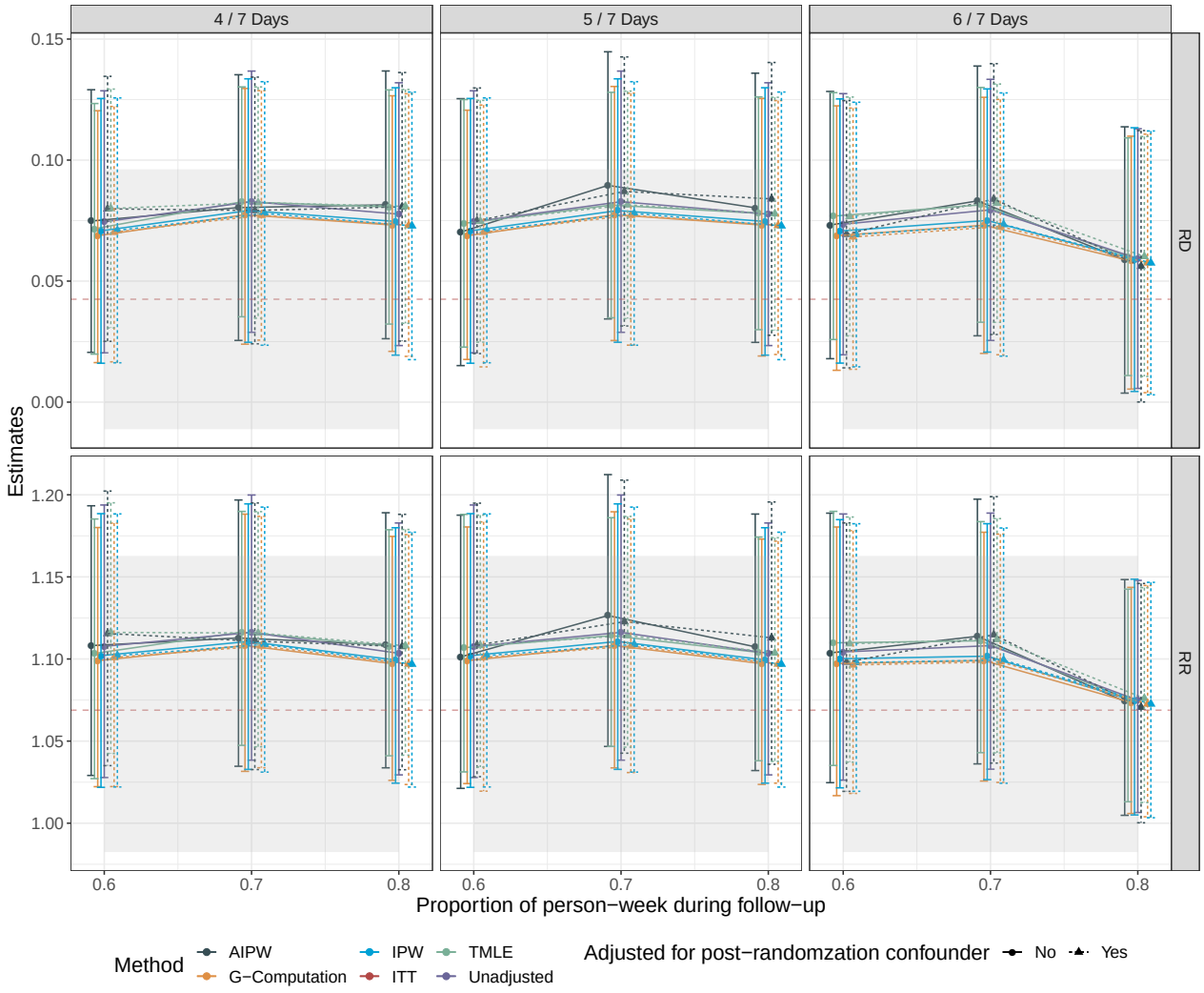
C.1 Sensitivity analyses of the per-protocol effects of low-dose aspirin on hCG-detected pregnancy

Table 12: Sensitivity analyses of the effects of low-dose aspirin on hCG-detected pregnancy among women adhered to the assigned treatment: 5/7 pills per week over at least 80% person-week of follow-up using different estimation methods

		RD				RR			
Method	Machine Learning	Est.	SE	LCL	UCL	Est.	SE	LCL	UCL
Intention-to-treat	No	0.043	0.027	-0.011	0.096	1.069	0.043	0.982	1.163
Per-protocol analysis adjusted for baseline covariates									
AIPW	Yes	0.080	0.028	0.025	0.136	1.107	0.036	1.032	1.188
TMLE	Yes	0.078	0.025	0.030	0.126	1.104	0.031	1.038	1.174
G-computation	No	0.073	0.027	0.019	0.125	1.097	0.035	1.024	1.173
IPW	No	0.075	0.028	0.019	0.130	1.099	0.036	1.024	1.180
Per-protocol analysis adjusted for baseline covariates and post-randomization confounders									
AIPW	Yes	0.084	0.029	0.028	0.140	1.113	0.037	1.036	1.196
TMLE	Yes	0.078	0.025	0.030	0.126	1.104	0.031	1.038	1.174
G-computation	No	0.073	0.027	0.020	0.125	1.097	0.034	1.024	1.172
IPW	No	0.073	0.028	0.018	0.128	1.097	0.036	1.022	1.177
Unadjusted per-protocol analysis									
Unadjusted	No	0.078	0.028	0.023	0.132	1.103	0.035	1.029	1.183

* Adjusted for unusual bleeding [$\geq 1/7days(20\%)$ per week over $\geq 50\%$ person-week] and nausea and/or vomiting [$\geq 1/7days(20\%)$ per week over $\geq 20\%$ person-week]

Figure 20: Sensitivity Analyses of the Effects of low-dose aspirin on hCG conception using different adherence levels and estimation methods



C.2 R code for per-protocol effect estimation

Setup

Load necessary packages and dataset

```
packages <- c("tidyverse", "AIPW", "SuperLearner",
              "earth","ranger","xgboost")

for (package in packages) {
  if (!require(package, character.only=T, quietly=T)) {
    install.packages(package,repos='http://lib.stat.cmu.edu/R/CRAN')
  }
}

#read dataset
eager_analysis <- read_csv("eager_pp_df_20200615.csv")
```

Intention-to-treat

```
# unadjusted estimates
get_all_est <- function(x){
  mat <- as.matrix(x)
  p1 <- mat[2,2]/sum(mat[2,])
  p0 <- mat[1,2]/sum(mat[1,])
  res <- data.frame(
    #Est./ SE
    RD = c(p1-p0,
           sqrt(p1*(1-p1)/sum(mat[2,])+p0*(1-p0)/sum(mat[1,]))),
    logRR = c( log(p1/p0),
               sqrt((1-p1)/mat[2,2]+(1-p0)/mat[1,2])),
    logOR = c( log((p1/(1-p1))/(p0/(1-p0))),
               sqrt(1/mat[2,2]+1/mat[1,2]+1/mat[1,1]+1/mat[2,1]))
  ) %>%
    rbind(., .[1,] - 1.96 * .[2,], .[1,] + 1.96 * .[2,]) %>%
    bind_cols(N = sum(mat),
              Param = c("Est.","SE","LCL","UCL"),
              Method = "Unadj")
  return(res)
}
#ITT estimates
itt <- get_all_est(table(eager_analysis$treatment,eager_analysis$conception))
itt
```

Per-protocol

Machine Learning + AIPW Adjusting for Baseline Covariates

```
# Set seeds
set.seed(123)
# Define learners for stacking machine learning via SuperLearner
earth_learner   <- create.Learner("SL.earth",tune=list(degree=c(2,3)))
ranger_learner  <- create.Learner("SL.ranger", tune=list(min.node.size = c(30),
                                                         num.trees=c(500),
                                                         max.depth=c(2,3)))
xgboost_learner <- create.Learner("SL.xgboost", tune=list(minobspernode = c(30),
```

```
                                                        ntrees=c(500),
                                                        max_depth=c(2,3),
                                                        subsample=c(1)))

sl.lib <- c("SL.glm", "SL.glm.interaction",
            earth_learner$names,ranger_learner$names, xgboost_learner$names)

# Subset dataset to those adhered to the protocol
df <- eager_analysis %>% filter(weeks_0.7_pt_0.8==1)

# select baseline covariates
bl_cov <- df %>%
  select(income_1:hsCRP) %>%
  as.data.frame()

# AIPW estimation via AIPW package
aipw_fit1 <- AIPW$new(Y=df$conception,
                      A=df$treatment,
                      W=bl_cov,
                      g.SL.library = sl.lib,
                      Q.SL.library = sl.lib,
                      k_split = 10,
                      verbose = TRUE)$
  fit()$
  summary(g.bound=0.025)$
  # check positivity
  plot.p_score()
```

Machine Learning + AIPW Adjusting for Baseline Covariates + Post-randomization Confounders

```
# select post-randomization confounders
postRand_confounder <- df %>%
  select(bleed_0.2_pt_0.5,nausea_0.2_pt_0.2) %>%
  as.data.frame()

# AIPW estimator via AIPW package
aipw_fit2 <- AIPW$new(Y=df$conception,
                      A=df$treatment,
                      W=cbind(bl_cov,postRand_confounder),
                      g.SL.library = sl.lib,
                      Q.SL.library = sl.lib,
                      k_split = 10,
                      verbose = TRUE)$
  fit()$
  summary(g.bound=0.025)$
  # check positivity
  plot.p_score()
```

Bibliography

[1] Allen J. Wilcox, Donna Day Baird, and Clarice R. Weinberg. Time of Implantation of the Conceptus and Loss of Pregnancy. *N Engl J Med*, 340(23):1796–1799, 1999. Number: 23.

[2] Maria C. Magnus, Allen J. Wilcox, Nils-Halvdan Morken, Clarice R. Weinberg, and Siri E. Håberg. Role of maternal age and pregnancy history in risk of miscarriage: prospective register based study. *BMJ*, 364, March 2019.

[3] K. J. Sapra, A. C. McLain, J. M. Maisog, R. Sundaram, and G. M. Buck Louis. Successive time to pregnancy among women experiencing pregnancy loss. *Human Reproduction (Oxford, England)*, 29(11):2553–2559, November 2014.

[4] J. A. Martin and M. J. K. Osterman. Describing the increase in preterm births in the United States, 2014-2016. NCHS Data Brief. 2018;(312): 1–8. Technical report.

[5] Lauren M. Rossen, Katherine A. Ahrens, and Amy M. Branum. Trends in Risk of Pregnancy Loss Among US Women, 1990-2011. *Paediatric and Perinatal Epidemiology*, 32(1):19–29, January 2018. Number: 1.

[6] M. Di Nisio, L. Peters, and S. Middeldorp. Anticoagulants for the treatment of recurrent pregnancy loss in women without antiphospholipid syndrome. *Cochrane Database Syst Rev*, (2):Cd004734, 2005.

[7] S. Kaandorp, M. Di Nisio, M. Goddijn, and S. Middeldorp. Aspirin or anticoagulants for treating recurrent miscarriage in women without antiphospholipid syndrome. *Cochrane Database Syst Rev*, (1):Cd004734, 2009.

[8] S. P. Kaandorp, M. Goddijn, J. A. Post, B. A. Hutten, H. R. Verhoeve, K. Hamulyak, B. W. Mol, N. Folkeringa, M. Nahuis, D. N. Papatsonis, H. R. Buller, F. Veen, and S. Middeldorp. Aspirin plus heparin or aspirin alone in women with recurrent miscarriage. *N Engl J Med*, 362(17):1586–96, 2010.

[9] J. C. Gris, E. Mercier, I. Quere, G. Lavigne-Lissalde, E. Cochery-Nouvellon, M. Hoffet, S. Ripart-Neveu, M. L. Tailland, M. Dauzat, and P. Mares. Low-molecular-weight heparin versus low-dose aspirin in women with one fetal loss and a constitutional thrombophilic disorder. *Blood*, 103(10):3695–9, 2004.

[10] Matthew K Hoffman, Shivaprasad S Goudar, Bhalachandra S Kodkany, Mrityunjay Metgud, Manjunath Somannavar, Jean Okitawutshu, Adrien Lokangaka, Antoinette Tshefu, Carl L Bose, Abigail Mwapule, Musaku Mwenechanya, Elwyn Chomba, Waldemar A Carlo, Javier Chicuy, Lester Figueroa, Ana Garces, Nancy F Krebs, Saleem Jessani, Farnaz Zehra, Sarah Saleem, Robert L Goldenberg, Kunal Kurhe, Prabir Das, Archana Patel, Patricia L Hibberd, Emmah Achieng, Paul Nyongesa, Fabian Esamai, Edward A Liechty, Norman Goco, Jennifer Hemingway-Foday, Janet Moore, Tracy L Nolen, Elizabeth M McClure, Marion Koso-Thomas, Menachem Miodovnik, R Silver, Richard J Derman, Emmah Achieng, Melissa Bauserman, Carl Bose, Sherri Bucher, Waldemar Carlo, Umesh S Charantimath, Javier Chicuy, Elwyn Chomba, Prabir Das, Richard Derman, Fabian Esamai, Lester Figueroa, Ms Ganachari, Ana Garces, Noman Goco, Robert Goldenberg, Shivaprasad Goudar, Jennifer Hemingway-Foday, Patricia Hibberd, Matthew Hoffman, Narayan V Honnungar, Saleem Jessani, Avinash Kavi, Bhalachandra Kodkany, Marion Koso-Thomas, Nancy Krebs, Yogesh Kumar Shashikanth, Kunal Kurhe, Edward Liechty, Adrien Lokangaka, Emily MacGuire, Ashalata A Mallapur, Elizabeth McClure, Mrityunjay Metgud, Menachem Miodovnik, Janet Moore, Abigail Mwapule, Musaku Mwenechanya, Farnaz Naqvi, Seemab Naqvi, Robert Nathan, Tracy Nolen, Paul Nyongesa, Jean Okitawutshu, Suchita Parepalli, Archana Patel, Umesh Y Ramadurg, Sarah Saleem, Robert Silver, Manjunath Somannavar, Zahid Soomro, Antoinette Tshefu, Sunil S Vernekar, Dennis Wallace, and Farnaz Zehra. Low-dose aspirin for the prevention of preterm delivery in nulliparous women with a singleton pregnancy (ASPIRIN): a randomised, double-blind, placebo-controlled trial. *The Lancet*, 395(10220):285–293, January 2020.

[11] S. Roberge, K. H. Nicolaides, S. Demers, P. Villa, and E. Bujold. Prevention of perinatal death and adverse perinatal outcome using low-dose aspirin: a meta-analysis. *Ultrasound in Obstetrics & Gynecology*, 41(5):491–499, 2013. _eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/uog.12421.

[12] Enrique F. Schisterman, Robert M. Silver, Neil J. Perkins, Sunni L. Mumford, Brian W. Whitcomb, Joseph B. Stanford, Laurie L. Lesher, David Faraggi, Jean Wactawski-Wende, Richard W. Browne, Janet M. Townsend, Mark White, Anne M. Lynch, and Noya Galai. A Randomised Trial to Evaluate the Effects of Low-dose Aspirin in Gestation and Reproduction: Design and Baseline Characteristics: EAGeR trial design. *Paediatric and Perinatal Epidemiology*, 27(6):598–609, November 2013. Number: 6.

[13] Enrique F Schisterman, Robert M Silver, Laurie L Lesher, David Faraggi, Jean Wactawski-Wende, Janet M Townsend, Anne M Lynch, Neil J Perkins, Sunni L Mumford, and Noya Galai. Preconception low-dose aspirin and pregnancy outcomes: results from the EAGeR randomised trial. *The Lancet*, 384(9937):29–36, July 2014. Number: 9937.

[14] Enrique F. Schisterman, Sunni L. Mumford, Karen C. Schliep, Lindsey A. Sjaarda, Joseph B. Stanford, Laurie L. Lesher, Jean Wactawski-Wende, Anne M. Lynch, Janet M. Townsend, Neil J. Perkins, Shvetha M. Zarek, Michael Y. Tsai, Zhen Chen, David Faraggi, Noya Galai, and Robert M. Silver. Preconception Low Dose Aspirin and Time to Pregnancy: Findings From the Effects of Aspirin in Gestation and Reproduction Randomized Trial. *The Journal of Clinical Endocrinology & Metabolism*, 100(5):1785–1791, May 2015. Number: 5.

[15] Ashley I. Naimi, Neil J. Perkins, Lindsey A. Sjaarda, Sunni L. Mumford, Robert W. Platt, Robert M. Silver, and Enrique F. Schisterman. The Effect of Preconception-Initiated Low-Dose Aspirin on Human Chorionic Gonadotropin–Detected Pregnancy, Pregnancy Loss, and Live Birth. *Annals of Internal Medicine*, January 2021. Publisher: American College of Physicians.

[16] Sandeep K. Gupta. Intention-to-treat concept: A review. *Perspectives in Clinical Research*, 2(3):109–112, 2011. Number: 3.

[17] Miguel A. Hernán, Sonia Hernández-Díaz, and James M. Robins. Randomized Trials Analyzed as Observational Studies. *Annals of Internal Medicine*, September 2013.

[18] Dan Sheng and Mimi Y. Kim. The effects of non-compliance on intent-to-treat analysis of equivalence trials. *Statistics in Medicine*, 25(7):1183–1199, April 2006. Number: 7.

[19] Miguel A Hernán and Sonia Hernández-Díaz. Beyond the intention-to-treat in comparative effectiveness research. *Clin Trials*, 9(1):48–55, 2012. Number: 1.

[20] Pan Wu, Wan Tang, Tian Chen, Hua He, Douglas Gunzler, and Xin M. Tu. Causal Inference: A Statistical Paradigm for Inferring Causality. In Hua He, Pan Wu, and Ding-Geng (Din) Chen, editors, *Statistical Causal Inferences and Their Applications in Public Health Research*, ICSA Book Series in Statistics, pages 3–25. Springer International Publishing, Cham, 2016.

[21] Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of Causal Effects Using Instrumental Variables. *J Am Stat Assoc*, 91(434):444–455, 1996.

[22] Constantine E. Frangakis and Donald B. Rubin. Principal Stratification in Causal Inference. *Biometrics*, 58(1):21–29, 2002. Number: 1.

[23] James M. Robins. CORRECTING FOR NON—COMPLIANCE IN RANDOMIZED TRIAIS USING STRUCI'URAL NESTED MEAN MODELS. 1994.

[24] M. A. Hernán and J. M. Robins. *Causal inference: What if*. Chapman & Hall/CRC, Boca Raton, 2020.

[25] JoAnn E. Manson, Chrisandra L. Shufelt, and James M. Robins. The Potential for Postrandomization Confounding in Randomized Clinical Trials. *JAMA*, 315(21):2273–2274, June 2016. Number: 21.

[26] Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, 210:2–21, August 2018.

[27] Ramón Estruch, Emilio Ros, Jordi Salas-Salvadó, Maria-Isabel Covas, Dolores Corella, Fernando Arós, Enrique Gómez-Gracia, Valentina Ruiz-Gutiérrez, Miquel Fiol, José Lapetra, Rosa M. Lamuela-Raventos, Lluís Serra-Majem, Xavier Pintó, Josep Basora, Miguel A. Muñoz, José V. Sorlí, J. Alfredo Martínez, Montserrat Fitó, Alfredo Gea, Miguel A. Hernán, and Miguel A. Martínez-González. Primary Prevention of Cardiovascular Disease with a Mediterranean Diet Supplemented with Extra-Virgin Olive Oil or Nuts. *New England Journal of Medicine*, 378(25):e34, June 2018. Number: 25.

[28] Sander Greenland, Judea Pearl, and James M. Robins. Causal Diagrams for Epidemiologic Research. *Epidemiology*, 10(1):37–48, 1999. Number: 1 Publisher: Lippincott Williams & Wilkins.

[29] J. M. Robins. Marginal Structural Models versus Structural Nested Models as Tools for Causal Inference. In ME Halloran and D Berry, editors, *Statistical Models in Epidemiology: The Environment and Clinical Trials*, volume 16, pages 95–134. Springer-Verlag, New York, 1999.

[30] J. M. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Diseases*, 40 Suppl 2:139S–161S, 1987.

[31] Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3(0):96–146, 2009. Number: 0.

[32] Sander Greenland. Quantifying Biases in Causal Models: Classical Confounding vs Collider-Stratification Bias:. *Epidemiology*, 14(3):300–306, May 2003.

[33] Tyler J. VanderWeele and Ilya Shpitser. On the definition of a confounder. *The Annals of Statistics*, 41(1):196–220, February 2013.

[34] Tyler J. VanderWeele. Principles of confounder selection. *European Journal of Epidemiology*, 34(3):211–219, March 2019.

[35] Denis Talbot and Victoria Kubuta Massamba. A descriptive review of variable selection methods in four epidemiologic journals: there is still room for improvement. *European Journal of Epidemiology*, 34(8):725–730, August 2019.

[36] Wei Liu, M. Alan Brookhart, Sebastian Schneeweiss, Xiaojuan Mi, and Soko Setoguchi. Implications of M Bias in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology*, 176(10):938–948, November 2012. Publisher: Oxford Academic.

[37] Peng Ding and Luke W. Miratrix. To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias. *Journal of Causal Inference*, 3(1):41–57, March 2015. Publisher: De Gruyter Section: Journal of Causal Inference.

[38] Judea Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann Publishers, 1988.

[39] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. Adaptive computation and machine learning. MIT Press, Cambridge, MA, 2009.

[40] David E. Heckerman and Bharat N. Nathwani. An evaluation of the diagnostic accuracy of Pathfinder. *Computers and Biomedical Research*, 25(1):56–74, February 1992. Number: 1.

[41] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, 2nd edition, 2009.

[42] Frederick Eberhardt. Introduction to the foundations of causal discovery. *International Journal of Data Science and Analytics*, 3(2):81–91, March 2017. Number: 2.

[43] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, 10, 2019.

[44] Sander Greenland, Judea Pearl, and JM Robins. Causal diagrams for epidemiological research. *Epidemiol*, 10(1):37–48, 1999. Number: 1.

[45] Alexander P Keil, Stephen J Mooney, Michele Jonsson Funk, Stephen R Cole, Jessie K Edwards, and Daniel Westreich. RESOLVING AN APPARENT PARADOX IN DOUBLY ROBUST ESTIMATORS. *Am J Epidemiol*, 187(4):891–892, April 2018. number: 4.

[46] Larry Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[47] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, February 2018. Publisher: Oxford Academic.

[48] Edward H. Kennedy and Sivaraman Balakrishnan. Discussion of "Data-driven confounder selection via Markov and Bayesian networks" by Jenny H\"aggstr\"om. *arXiv:1710.11566 [stat]*, October 2017. arXiv: 1710.11566.

[49] Ashley I. Naimi, Alan E. Mishler, and Edward H. Kennedy. Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms. *American Journal of Epidemiology*, In Press, 2021. arXiv: 1711.07137.

[50] Paul N. Zivich and Alexander Breskin. Machine Learning for Causal Inference: On the Use of Cross-fit Estimators. *Epidemiology*, 32(3):393–401, May 2021.

[51] James M. Robins and Andrea Rotnitzky. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 90(429):122–129, 1995. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[52] Jeremy Coyle and Mark J. van der Laan. Targeted Bootstrap. In *Targeted Learning in Data Science*, pages 523–539. Springer International Publishing, Cham, 2018. Series Title: Springer Series in Statistics.

[53] Iván Díaz. Machine learning in the estimation of causal effects: targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, 21(2):353–358, April 2020. Publisher: Oxford Academic.

[54] Michael Lamm and Yiu-Fai Yung. Estimating causal effects from observational data with the CAUSALTRT procedure. In *Proceedings of the SAS Global Forum 2017 Conference. Cary, NC: SAS Institute Inc. http://support. sas. com/resources/papers/proceedings17/SAS0374-2017. pdf*, 2017.

[55] Bryan S. Graham, C. Campos De Xavier Pinto, and Daniel Egel. Inverse Probability Tilting Estimation of Average Treatment Effects in Stata. *The Stata Journal*, pages 1–16, 2011.

[56] Adam N. Glynn and Kevin M. Quinn. An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis*, 18(1):36–56, 2010. Number: 1.

[57] S. Roberge, K. H. Nicolaides, S. Demers, P. Villa, and E. Bujold. Prevention of perinatal death and adverse perinatal outcome using low-dose aspirin: a meta-analysis. *Ultrasound in Obstetrics & Gynecology*, 41(5):491–499, 2013. _eprint: https://obgyn.onlinelibrary.wiley.com/doi/pdf/10.1002/uog.12421.

[58] David Benkeser, Iván Díaz, Alex Luedtke, Jodi Segal, Daniel Scharfstein, and Michael Rosenblum. Improving precision and power in randomized trials for COVID-19 treatments using covariate adjustment, for binary, ordinal, and time-to-event outcomes. *Biometrics*, n/a(n/a), September 2020. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.13377.

[59] Sander Greenland and Babette Brumback. An overview of relations among causal modelling methods. *International Journal of Epidemiology*, 31(5):1030–1037, October 2002. Number: 5 Publisher: Oxford Academic.

[60] Peter W G Tennant, Eleanor J Murray, Kellyn F Arnold, Laurie Berrie, Matthew P Fox, Sarah C Gadd, Wendy J Harrison, Claire Keeble, Lynsie R Ranker, Johannes Textor, Georgia D Tomova, Mark S Gilthorpe, and George T H Ellison. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *International Journal of Epidemiology*, (dyaa213), December 2020.

[61] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, Cambridge, 2000.

[62] Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, December 2009. Number: 6A.

[63] P. Ding, T. J. Vanderweele, and J. M. Robins. Instrumental variables as bias amplifiers with general outcome and confounding. *Biometrika*, 104(2):291–302, June 2017. Publisher: Oxford Academic.

[64] Trang Quynh Nguyen, Allan Dafoe, and Elizabeth L. Ogburn. The Magnitude and Direction of Collider Bias for Binary Variables. *Epidemiologic Methods*, 8(1), March 2019. Publisher: De Gruyter Section: Epidemiologic Methods.

[65] Emilija Perković, Johannes Textor, and Markus Kalisch. Complete Graphical Characterization and Construction of Adjustment Sets in Markov Equivalence Classes of Ancestral Graphs. page 62.

[66] Clark Glymour and Richard Scheines. Causal modeling with the TETRAD program. page 27, 1986.

[67] Clark N. Scheines Richard Spirtes, Peter Glymour. *Causation, prediction, and search*. MIT Press, Cambridge, Mass., 1993.

[68] Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3, February 2016. Number: 1.

[69] Doris Entner, Patrik Hoyer, and Peter Spirtes. Data-driven covariate selection for nonparametric estimation of causal effects. In *Artificial Intelligence and Statistics*, pages 256–264. PMLR, 2013.

[70] Jenny Häggström. Data-driven confounder selection via Markov and Bayesian networks. *Biometrics*, 74(2):389–398, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12788.

[71] Vishesh Karwa, Aleksandra B. Slavković, and Eric T. Donnell. Causal inference in transportation safety studies: Comparison of potential outcomes and causal diagrams. *The Annals of Applied Statistics*, 5(2B):1428–1455, June 2011. Number: 2B.

[72] M Maria Glymour. Using causal diagrams to understand common problems in social epidemiology. In J. M. Oakes and J. S. Kaufman, editors, *Methods in Social Epidemiology*, pages 393–428. Jossey-Bass: San Francisco, CA, 2006.

[73] Jacqueline E. Rudolph, Jessie K. Edwards, Ashley I. Naimi, and Daniel J. Westreich. Simulation in Practice: The Balancing Intercept. *American Journal of Epidemiology*.

[74] Miguel A Hernán and James M Robins. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*, 60(7):578–586, 2006. Number: 7.

[75] Susan Gruber and Mark J van der Laan. tmle: An R Package for Targeted Maximum Likelihood Estimation. *Journal of Statistical Software*, 51(13):1–35, 2012. Number: 13.

[76] Ashley I. Naimi, Stephen R. Cole, and Edward H. Kennedy. An introduction to g methods. *International journal of epidemiology*, 46(2):756–762, 2017. ISBN: 0300-5771 Publisher: Oxford University Press.

[77] Yongqi Zhong, Edward H. Kennedy, Lisa M. Bodnar, and Ashley I. Naimi. AIPW: An R Package for Augmented Inverse Probability Weighted Estimation of Average Causal Effects. *American Journal of Epidemiology*, In Press, 2021.

[78] Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing Bayesian network structure learning algorithm. *Mach Learn*, page 48, 2006.

[79] Bryan Keller. Variable Selection for Causal Effect Estimation: Nonparametric Conditional Independence Testing With Random Forests. *Journal of Educational and Behavioral Statistics*, 45(2):119–142, April 2020. Publisher: American Educational Research Association.

[80] Marco Scutari. Learning Bayesian Networks with the bnlearn R Package. *arXiv:0908.3817 [stat]*, August 2009. arXiv: 0908.3817.

[81] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[82] Mikko Koivisto. Advances in exact Bayesian structure discovery in Bayesian networks. *arXiv:1206.6828 [cs, stat]*, June 2012. arXiv: 1206.6828.

[83] Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. Challenges and Opportunities with Causal Discovery Algorithms: Application to Alzheimer's Pathophysiology. *Scientific Reports*, 10(1):2975, February 2020. Number: 1 Publisher: Nature Publishing Group.

[84] Senol Isci, Cengizhan Ozturk, Jon Jones, and Hasan H. Otu. Pathway analysis of high-throughput biological data within a Bayesian network framework. *Bioinformatics*, 27(12):1667–1674, June 2011.

[85] Swathi P. Iyer, Izhak Shafran, David Grayson, Kathleen Gates, Joel T. Nigg, and Damien A. Fair. Inferring functional connectivity in MRI using Bayesian network structure learning with a modified PC algorithm. *NeuroImage*, 75:165–175, July 2013.

[86] Chengwei Su, Angeline Andrew, Margaret R. Karagas, and Mark E. Borsuk. Using Bayesian networks to discover relations between genes, environment, and disease. *BioData Mining*, 6(1):6, March 2013.

[87] Yi-Chun Chen, Tim A. Wheeler, and Mykel J. Kochenderfer. Learning Discrete Bayesian Networks from Continuous Data. *Journal of Artificial Intelligence Research*, 59:103–132, June 2017.

[88] Jiaying Gu, Fei Fu, and Qing Zhou. Penalized estimation of directed acyclic graphs from discrete data. *Statistics and Computing*, 29(1):161–176, January 2019. Number: 1.

[89] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, February 2002. Publisher: National Academy of Sciences Section: Colloquium Paper.

[90] Lixia Zhang, Leonardo O. Rodrigues, Niven R. Narain, and Viatcheslav R. Akmaev. bAIcis: A Novel Bayesian Network Structural Learning Algorithm and Its Comprehensive Performance Evaluation Against Open-Source Software. *Journal of Computational Biology*, September 2019.

[91] RICHARD Bellman. Dynamic programming, princeton univ. *Prese} Princeton, 19S7*, 1957.

[92] Antonio Salmerón, Rafael Rumí, Helge Langseth, Thomas D. Nielsen, and Anders L. Madsen. A Review of Inference Algorithms for Hybrid Bayesian Networks. *J. Artif. Intell. Res.*, 62:799–828, 2018.

[93] Sebastian Schneeweiss. Automated data-adaptive analytics for electronic healthcare data to study causal treatment effects. *Clinical Epidemiology*, Volume 10:771–788, July 2018.

[94] Konstantina Biza, Ioannis Tsamardinos, and Sofia Triantafillou. Tuning causal discovery algorithms. *Probabilistic Graphical Models*, 2020.

[95] James M. Robins and Larry Wasserman. On the impossibility of inferring causation from association without background knowledge. *Computation, causation, and discovery*, 1999:305–21, 1999. Publisher: Menlo Park, CA, Cambridge, MA: AAAI Press/The MIT Press.

[96] James M. Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003. ISBN: 1464-3510 Publisher: Oxford University Press.

[97] Daniel Westreich, Justin Lessler, and Michele Jonsson Funk. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833, August 2010. Number: 8.

[98] Brian K. Lee, Justin Lessler, and Elizabeth A. Stuart. Improving propensity score weighting using machine learning. *Stat Med*, 29(3):337–346, 2010. Number: 3.

[99] Ariel Linden and Paul R. Yarnold. Combining machine learning and matching techniques to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, 22(6):868–874, 2016. Number: 6 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jep.12592.

[100] Min Lu, Saad Sadiq, Daniel J. Feaster, and Hemant Ishwaran. Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods. *Journal of Computational and Graphical Statistics*, 27(1):209–219, January 2018. Number: 1 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10618600.2017.1356325.

[101] Tony Blakely, John Lynch, Koen Simons, Rebecca Bentley, and Sherri Rose. Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference. *International Journal of Epidemiology*, June 2020.

[102] Edward H. Kennedy, Sivaraman Balakrishnan, and Larry A. Wasserman. Discussion of "On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning". *arXiv:2006.09613 [stat]*, June 2020. arXiv: 2006.09613.

[103] Edward H Kennedy. Semiparametric theory and empirical processes in causal inference. In Hua He, Pan Wu, and Ding-Geng (Din) Chen, editors, *Statistical Causal Inferences and Their Applications in Public Health Research*. Springer International, Switzerland, 2016.

[104] Sherri Rose and Mark J Laan. *Targeted learning: causal inference for observational and experimental data*. Springer, New York, NY, 2011.

[105] Megan S. Schuler and Sherri Rose. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. *American Journal of Epidemiology*, 185(1):65, 2017. Number: 1.

[106] Jeremy R Coyle and Nima S Hejazi. tlverse/tmle3, August 2020. original-date: 2017-10-20T18:47:10Z.

[107] Edward H Kennedy. ehkennedy/npcausal, August 2020. original-date: 2017-05-18T02:08:13Z.

[108] Klaus K. Holst. kkholst/targeted, May 2020. original-date: 2020-04-13T09:00:18Z.

[109] Paul Zivich. pzivich/zEpid, July 2020. original-date: 2017-10-10T11:47:47Z.

[110] Bryan S. Graham, Cristine Campos de Xavier Pinto, and Daniel Egel. Inverse probability tilting for moment condition models with missing data. *The Review of*

Economic Studies, 79(3):1053–1079, 2012. ISBN: 1467-937X Publisher: Oxford University Press.

[111] Thomas S. Richardson, James M. Robins, and Linbo Wang. On Modeling and Estimation for the Relative Risk and Risk Difference. *Journal of the American Statistical Association*, 112(519):1121–1130, July 2017. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2016.1192546.

[112] Peter C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3):399–424, May 2011.

[113] Fan Li, Kari Lock Morgan, and Alan M. Zaslavsky. Balancing Covariates via Propensity Score Weighting. *Journal of the American Statistical Association*, 113(521):390–400, January 2018.

[114] J. M. Robins, S. D. Mark, and W. K. Newey. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(2):479–95, 1992. Number: 2.

[115] Shaun R. Seaman and Stijn Vansteelandt. Introduction to Double Robust Methods for Incomplete Data. *Statistical Science*, 33(2):184–197, May 2018. Number: 2.

[116] Heejung Bang and James M. Robins. Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962–973, 2005. Number: 4.

[117] Michele Jonsson-Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, and Marie Davidian. Doubly Robust Estimation of Causal Effects. *Am J Epidemiol*, 173(7):761–767, 2011. Number: 7.

[118] Aaron Fisher and Edward H. Kennedy. Visually Communicating and Teaching Intuition for Influence Functions. *The American Statistician*, pages 1–11, February 2020.

[119] Hadley Wickham. *Advanced r*. CRC press, 2019.

[120] Winston Chang. r-lib/R6, August 2020. original-date: 2014-05-07T04:33:54Z.

[121] Ashley I Naimi and Laura B Balzer. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*, 33(5):459–464, 2018. Number: 5.

[122] Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):Article 25, 2007. Number: 1.

[123] Trevor J. Hastie and Robert J. Tibshirani. *Generalized additive models*, volume 43. CRC press, 1990.

[124] Jerome H. Friedman. Multivariate Adaptive Regression Splines. *Annals of Statistics*, 19(1):1–67, March 1991. Publisher: Institute of Mathematical Statistics.

[125] Marvin N. Wright and Andreas Ziegler. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, 77(1), 2017. arXiv: 1508.04409.

[126] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, August 2016. Association for Computing Machinery.

[127] Wenjing Zheng and Mark J. Van Der Laan. Asymptotic Theory for Cross-validated Targeted Maximum Likelihood Estimation. *U.C. Berkeley Division of Biostatistics Working Paper Series*, November 2010.

[128] E. H. Kennedy, A. Sjölander, and D. S. Small. Semiparametric causal inference in matched cohort studies. *Biometrika*, 102(3):739–746, September 2015.

[129] Tim P. Morris, Ian R. White, and Michael J. Crowther. Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, 38(11):2074–2102, 2019. Number: 11 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8086.

[130] Eric Polley, Erin LeDell, Chris Kennedy, Sam Lendle, and Mark van der Laan. SuperLearner: Super Learner Prediction, December 2019.

[131] Jeremy R Coyle, Nima S Hejazi, Ivana Malenica, and Oleg Sofrygin. sl3: Modern Super Learning with Pipelines, March 2020.

[132] Anastasios A. Tsiatis, Marie Davidian, Min Zhang, and Xiaomin Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27(23):4658–4677, 2008. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.3113.

[133] Elizabeth Colantuoni and Michael Rosenblum. Leveraging prognostic baseline variables to gain precision in randomized trials. *Statistics in Medicine*, 34(18):2602–2617, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.6507.

[134] Iván Díaz, Elizabeth Colantuoni, and Michael Rosenblum. Enhanced precision in the analysis of randomized trials with ordinal outcomes. *Biometrics*, 72(2):422–431, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12450.

[135] Edward H. Kennedy, Sivaraman Balakrishnan, and Max G'Sell. Sharp instruments for classifying compliers and generalizing causal effects. *arXiv:1801.03635 [stat]*, May 2019. arXiv: 1801.03635.

[136] Whitney K. Newey and James R. Robins. Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation. *arXiv:1801.09138 [math, stat]*, January 2018. arXiv: 1801.09138.

[137] Zhiqiang Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682, 2010. Number: 3.

[138] Miguel A. Hernán and James M. Robins. Per-Protocol Analyses of Pragmatic Trials. *New England Journal of Medicine*, 377(14):1391–1398, October 2017.

[139] Lauren E. Cain and Stephen R. Cole. Inverse probability-of-censoring weights for the correction of time-varying noncompliance in the effect of randomized highly active antiretroviral therapy on incident AIDS or death. *Statistics in Medicine*, 28(12):1725–1738, May 2009.

[140] Sara Lodi, Shweta Sharma, Jens D. Lundgren, Andrew N. Phillips, Stephen R. Cole, Roger Logan, Brian K. Agan, Abdel Babiker, Hartwig Klinker, Haitao Chu, Matthew Law, James D. Neaton, and Miguel A. Hernán. The per-protocol effect of immediate versus deferred antiretroviral therapy initiation. *AIDS*, 30(17):2659–2663, November 2016.

[141] Sara Lodi, Matthew Freiberg, Natalia Gnatienko, Elena Blokhina, Tatiana Yaroslavtseva, Evgeny Krupitsky, Eleanor Murray, Jeffrey H. Samet, and Debbie M. Cheng. Per-protocol analysis of the ZINC trial for HIV disease among alcohol users. *Trials*, 22(1):1–7, December 2021. Number: 1 Publisher: BioMed Central.

[142] Pamela M. Murnane, Elizabeth R. Brown, Deborah Donnell, R. Yates Coley, Nelly Mugo, Andrew Mujugira, Connie Celum, and Jared M. Baeten. Estimating Efficacy in a Randomized Trial With Product Nonadherence: Application of Multiple Methods to a Trial of Preexposure Prophylaxis for HIV Prevention. *American Journal of Epidemiology*, 182(10):848–856, November 2015.

[143] Eleanor J. Murray and Miguel A. Hernán. Improved adherence adjustment in the Coronary Drug Project. *Trials*, 19(1):158, December 2018.

[144] Anke Neumann, Géric Maura, Alain Weill, Philippe Ricordeau, François Alla, and Hubert Allemand. Comparative effectiveness of rosuvastatin versus simvastatin in primary prevention among new users: a cohort study in the French national health insurance database. *Pharmacoepidemiology and Drug Safety*, 23(3):240–250, March 2014.

[145] Sengwee Toh, Sonia Hernández-Díaz, Roger Logan, James M. Robins, and Miguel A. Hernán. Estimating Absolute Risks in the Presence of Nonadherence: An Application to a Follow-up Study With Baseline Randomization. *Epidemiology*, 21(4):528–539, July 2010.

[146] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2009.

[147] Mark P. Little, Pauline Brocard, Paul Elliott, and Philip J. Steer. Hemoglobin concentration in pregnancy and perinatal mortality: a London-based cohort study. *American Journal of Obstetrics and Gynecology*, 193(1):220–226, July 2005.

[148] Leo Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001. Number: 1.

[149] Sherri Rose and Mark J van der Laan. *Targeted learning: causal inference for observational and experimental data*. Springer, New York, NY, 2011.

[150] Lawrence Carin and Michael J. Pencina. On Deep Learning for Medical Image Analysis. *JAMA*, 320(11):1192–1193, September 2018.

[151] Shuojia Wang, Jyotishman Pathak, and Yiye Zhang. Using Electronic Health Records and Machine Learning to Predict Postpartum Depression. *MEDINFO 2019: Health and Wellbeing e-Networks for All*, pages 888–892, 2019. Publisher: IOS Press.

[152] Rohan Khera, Julian Haimovich, Nathan C. Hurley, Robert McNamara, John A. Spertus, Nihar Desai, John S. Rumsfeld, Frederick A. Masoudi, Chenxi Huang, Sharon-Lise Normand, Bobak J. Mortazavi, and Harlan M. Krumholz. Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction. *JAMA Cardiology*, 6(6):633, June 2021.

[153] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Random forests. In *The elements of statistical learning*, pages 587–604. Springer, 2009.

[154] Adele Cutler, D. Richard Cutler, and John R. Stevens. Random Forests. In Cha Zhang and Yunqian Ma, editors, *Ensemble Machine Learning: Methods and Applications*, pages 157–175. Springer US, Boston, MA, 2012.

[155] Graham Dunn and M. Lovrić. Complier-average causal effect (CACE) estimation. In *International encyclopedia of statistical science*. Springer Nature, 2011.

[156] Awtry Eric H. and Loscalzo Joseph. Aspirin. *Circulation*, 101(10):1206–1218, March 2000. Publisher: American Heart Association.

[157] Iosief Abraha and Alessandro Montedori. Modified intention to treat reporting in randomised controlled trials: systematic review. *BMJ (Clinical research ed.)*, 340:c2697, June 2010.

[158] Alessandro Montedori, Maria Isabella Bonacini, Giovanni Casazza, Maria Laura Luchetta, Piergiorgio Duca, Francesco Cozzolino, and Iosief Abraha. Modified versus standard intention-to-treat reporting: Are there differences in methodological quality, sponsorship, and findings in randomized trials? A cross-sectional study. *Trials*, 12(1):1–9, December 2011. Number: 1 Publisher: BioMed Central.

[159] Jessica M. Franklin, Sebastian Schneeweiss, Jennifer M. Polinski, and Jeremy A. Rassen. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics & Data Analysis*, 72:219–226, April 2014.

[160] Xun Zheng, Chen Dan, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. Learning Sparse Nonparametric DAGs. *arXiv:1909.13189 [cs, stat]*, September 2019. arXiv: 1909.13189.

[161] Samuel D. Lendle, Joshua Schwab, Maya L. Petersen, and Mark J. van der Laan. ltmle: An R Package Implementing Targeted Minimum Loss-Based Estimation for Longitudinal Data. *Journal of Statistical Software*, 81(1):1–21, October 2017. Number: 1.

[162] Andrea Rotnitzky and James Robins. Inverse probability weighted estimation in survival analysis. page 13.

[163] James M Robins and Miguel Á Hernán. Estimation of the Causal Effects of Time-Varying Exposures. In G Fitzmaurice, M. Davidian, G Verbeke, and G Molenberghs, editors, *Advances in Longitudinal Data Analysis*, pages 553–599. Chapman & Hall, Boca Raton, FL, 2009.

[164] Miguel A. Hernán and Tyler J. VanderWeele. Compound Treatments and Transportability of Causal Inference. *Epidemiology (Cambridge, Mass.)*, 22(3):368–377, May 2011.

[165] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, July 2016. Publisher: National Academy of Sciences Section: Colloquium Paper.

[166] Daniel Westreich, Jessie K Edwards, Catherine R Lesko, Elizabeth Stuart, and Stephen R Cole. Transportability of Trial Results Using Inverse Odds of Sampling Weights. *American Journal of Epidemiology*, 186(8):1010–1014, October 2017.

[167] Catherine R. Lesko, Lisa P. Jacobson, Keri N. Althoff, Alison G. Abraham, Stephen J. Gange, Richard D. Moore, Sharada Modur, and Bryan Lau. Collaborative, pooled and harmonized study designs for epidemiologic research: challenges and opportunities. *International Journal of Epidemiology*, 47(2):654–668, April 2018.

[168] Anjani Chandra, Gladys M. Martinez, William D. Mosher, Joyce C. Abma, and Jo Jones. Fertility, family planning, and reproductive health of US women; data from the 2002 National Survey of Family Growth. 2005.

[169] Lepkowski Jm, Mosher Wd, Davis Ke, Groves Rm, and Van Hoewyk J. The 2006-2010 National Survey of Family Growth: sample design and analysis of a continuous survey. *Vital and Health statistics. Series 2, Data Evaluation and Methods Research*, (150):1–36, June 2010.

[170] Lydia J. Leon, Fergus P. McCarthy, Kenan Direk, Arturo Gonzalez-Izquierdo, David Prieto-Merino, Juan P. Casas, and Lucy Chappell. Preeclampsia and Cardiovascular Disease in a Large UK Pregnancy Cohort of Linked Electronic Health Records. *Circulation*, 140(13):1050–1060, September 2019. Publisher: American Heart Association.

[171] Hannah R. Simons and Julia E. Kohn. Examining Temporal Trends in Documentation of Pregnancy Intentions in Family Planning Health Centers Using Electronic Health Records. *Maternal and Child Health Journal*, 23(1):47–53, January 2019.

[172] Nitzan Shalom Artzi, Smadar Shilo, Eran Hadar, Hagai Rossman, Shiri Barbash-Hazan, Avi Ben-Haroush, Ran D. Balicer, Becca Feldman, Arnon Wiznitzer, and Eran Segal. Prediction of gestational diabetes based on nationwide electronic health records. *Nature Medicine*, 26(1):71–76, January 2020. Number: 1 Publisher: Nature Publishing Group.