MISSING DATA AND MEASUREMENT ERROR: ANALYTIC APPROACHES FOR OBSERVATIONAL STUDIES WITH EXAMPLES FROM PREGNANCY OUTCOMES RESEARCH

Rachael K. Ross

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Epidemiology in the Gillings School of Global Public Health.

Chapel Hill
2023

Approved by

Stephen R. Cole

Julie L. Daniels

Jessie K. Edwards

Jeffrey S.A. Stringer

Daniel Westreich

ABSTRACT

Rachael K. Ross: Missing Data and Measurement Error: Analytic Approaches for Observational Studies With Examples from Pregnancy Outcomes Research
(Under the direction of Stephen Cole)

Observational research to improve pregnancy outcomes faces methodological challenges that limit accurate and actionable evidence. The objective of this dissertation was to examine analytic approaches to address the common challenges of measurement error and missing data.

In Aim 1, we developed and validated standardization (g-computation) estimators that leverage external validation data to account for outcome misclassification. To account for measurement error, external validation data are used to estimate misclassification probabilities (i.e., sensitivity and specificity). When the validation data are external, the estimated misclassification probabilities may need to be transported from the validation to the target population. If there are variables related to misclassification whose distribution differ between the validation and target, these probabilities are not immediately transportable. We introduce estimators that account for these variables in order to transport. For estimation, we used M-estimation. In simulation, these estimators were unbiased when assumptions were met and confidence intervals had appropriate coverage. We used these estimators in an applied example to estimate the effect of maternal HIV infection on preterm birth. Estimates accounting for outcome misclassification were notably different from the naïve analysis.

In Aim 2, we illustrated implementation and examined the performance of a novel weighted estimator to address nonmonotone missingness. In simulation, we compared performance to complete case analysis and multiple imputation by chained equations. Regardless of the missing data approach, we used weighting to address confounding. When complete case analysis was biased, weighting and imputation were unbiased, except when data were missing not at random. Imputation was more precise as

sample size and percent exposed declined; otherwise imputation and weighting were similarly precise. Imputation was less computationally efficient than weighting. We used these estimators in an applied example to estimate the effect of maternal anemia on preterm birth risk, where estimates were similar across approaches.

Measurement error and missing data are often overlooked yet they can produce substantial bias. This dissertation examined novel analytic tools to address these issues. Our work makes these tools accessible to other epidemiologists in order to advance research to improve pregnancy outcomes and public health more broadly.

To Sanders

ACKNOWLEDGMENTS

I would like to acknowledge a number of people who have supported and encouraged me during the journey of writing this dissertation. First and foremost, I would like to thank my advisor, Steve Cole, for his guidance, expertise, and encouragement throughout my graduate studies. I learned more than I ever imagined I would when I started this program, in large part because of working with you. Your support and mentorship have been invaluable in shaping my research and academic journey.

I would also like to extend my gratitude to my committee members, Jess Edwards, Daniel Westreich, Julie Daniels, and Jeff Stringer, for their time, feedback, and contributions to my research. Your insightful comments and constructive criticisms have helped me to refine my work and challenged me to think more critically about my research questions.

I would like to thank my peers and friends, in particular Emilie Duchesneau, Leah Sadinski, Audrey Renson, and Karen Diepstra, for sharing their knowledge and experiences, as well and many dinners, bike rides, and walks around Carrboro. Your presence and contributions have enriched and enlivened the past five years.

Finally, I would like to thank my partner Chris for his love and doing the laundry and grocery shopping, and ensuring we get a healthy dose of nature every week. Despite your hesitations about moving to the Southeast, you embraced North Carolina and have brought so much fun and adventure to our life here. You and Sanders helped me maintain a balance that made pursuing this degree such an enjoyable endeavor. I'm grateful for the life we have made together.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Avg. ModSE	Average model standard error
Avg. SE	Average estimated standard error
CBE	Constrained Bayesian estimator
CC	Complete case/complete case analysis
CI	Confidence interval
ESE	Empirical standard error
HIV	Human Immunodeficiency Virus
IPOP	Improving Pregnancy Outcomes with Progesterone
IQR	Interquartile range
JAGS	Just Another Gibbs Sampler
LMP	Last menstrual period
MAR	Missing at random
MCAR	Missing completely at random
MI	Multiple imputation
MNAR	Missing not at random
PTB	Preterm Birth
RMSE	Root mean squared error
SE	Standard error
Se	Sensitivity
SGA	Small for gestational age
Sp	Specificity
Spont. PTB	Spontaneous preterm birth
UMLE	Unconstrainted maximum likelihood estimator
ZAPPS	Zambia Preterm Birth Prevention Study
ZEPRS	Zambia Electronic Perinatal Record System

CHAPTER 1: INTRODUCTION

The need for effective interventions to prevent adverse pregnancy outcomes

Every year millions of pregnancies result in fetal death or severe long-term morbidity and highly effective interventions to prevent adverse outcomes remain elusive. Annually, there are more than 2.5 million stillbirths, 13.4 million preterm births, and 32 million babies born small for gestational age (SGA) in low- and middle-income countries.[1–5] Over one third of all livebirths in low- and middle-income countries are preterm or SGA.[4] Complications of preterm birth are the leading cause of neonatal death and the 2nd leading cause of death in children under 5 years old.[6–8] In addition to elevated mortality risk, children born preterm or SGA are at higher risk of short- and long-term effects on immunologic competence, visual impairment, neurodevelopmental functioning, chronic lung disease, and adult-onset chronic conditions.[2,9–13] These adverse birth outcomes are also associated with high financial, psychological, and social costs.[14]

Despite the incredibly high burden and severity of these adverse pregnancy outcomes, there are few effective evidence-based interventions to prevent them. Many trials for interventions aimed at preventing these outcomes have failed.[15,16] For preterm birth, for example, trials for nutritional and protein supplementation, and screening and treatment of infections have shown little benefit.[17] Just two interventions have strong evidence for the prevention of preterm birth and are recommended for low- and middle-income countries: administration of progesterone and smoking cessation programs.[17] However, a recent trial in Zambia found no effect of progesterone on the prevention preterm birth among HIV-infected people and smoking among pregnant people is uncommon in much of sub-Saharan Africa.[18,19] There are some effective evidence-based interventions for stillbirth and SGA such as protein-energy, micronutrient, and folic acid supplementation, and syphilis screening and treatment, but more effective

interventions are needed as well as research on implementation and integration of interventions into care.[20–24]

Pregnancy outcomes research using observational data is plagued with methodological challenges that hamper scientific advancement. Although clinical trials are often the preferred study design to produce evidence to inform intervention, trials can be prohibitively expensive and take a long time to implement. Further, results may not be generalizable due to the strict inclusion and exclusion criteria typical of trials.[25] By emulating a trial in study design,[26–28] observational data can be used to estimate the impact of hypothetical interventions by leveraging natural variation in the care and treatment of pregnant people, however research using observational data requires strong untestable assumptions (e.g., conditional exchangeability with positivity and causal consistency) that are met in expectation or by design in randomized trials.[29] Observational data sources are also subject to measurement error and missing data. While data from trials are not immune to these issues, they are often avoided or mitigated by strict use of protocols and greater resources. The challenges of using observational data for pregnancy outcomes research, including measurement error and missing data, are often ignored resulting in estimates that are uninformative.

Measurement error

Gestational age measurement error

Gestational age is used to define important pregnancy outcomes such as preterm birth (birth prior to 37 completed weeks of gestation). In research studies that enroll people during pregnancy, the exact date of conception is unknown so gestational age must be estimated. Commonly, gestational age is estimated from the date of the last menstrual period (LMP) as recalled by the pregnant person or from fetal biometry measured by ultrasound.[30] Dating by LMP has two sources of error: variation in a person's follicular-phase length (the time between LMP and ovulation) and recall.[31–39] The amount of error may be related to a number of factors associated with poor pregnancy outcomes such as maternal age, pregnancy history, and delayed start of prenatal care.[35–37,40] Gestational age dating by ultrasound is more accurate

than dating by LMP, particularly before 24 weeks gestation.[37,41–46] Unfortunately, due to lack of resources, ultrasound measurement is not universally available in low- and middle-income countries.

Measurement error can produce substantial and unpredictable bias, but it often goes unaddressed.[47–50] It has been shown that using LMP-based gestational age widens the distribution of gestational age at birth (i.e., gestational duration) compared to ultrasound measurement, resulting in an excess of apparent preterm births.[40,42] For example, in Zambia Preterm Birth Prevention Study, the proportion of births classified as preterm was 17.8% when estimated by LMP and 13.8% when estimated by ultrasound. Studies have consistently observed such overestimation of the proportion of preterm births by LMP compared to ultrasound[40,51–53] and these differences in estimates can have a substantial impact on public health planning.

The impact of outcome misclassification

When estimating the effect of an exposure on the risk of preterm birth, measurement error can produce bias towards or away from the null. Under specific conditions measurement error in the outcome will not produce bias or the bias will be in a predictable direction.[48,54(sec1.2),55,56] However, beyond simple and often unrealistic scenarios, it is challenging to predict how measurement error in the outcome will impact results, particularly so when error is differential or dependent.[55,57,58]

Analytic approaches can be used to address measurement error and produce more accurate estimates. These approaches include regression calibration, multiple imputation for measurement error, and maximum likelihood.[57,59–63] In some areas of epidemiologic research such as nutritional and occupational studies, these analytic approaches are more common, however they are rarely applied in pregnancy outcomes research. There are a number of potential barriers to uptake of these approaches: measurement error is not generally directly observable so it is easy to overlook; there is limited appreciation for the impact of bias from measurement error; there is a strong perception that the direction of bias is predictable; and many measurement error approaches are statistically complex making adoption challenging for epidemiologists with limited statistical training.[48] At present, the majority of research on adverse pregnancy outcomes from low- and middle-income countries suffers from bias due to gestational

age measurement error and without greater awareness of its impact and easy to implement approaches for correction, measurement error will continue to be a barrier to producing robust and accurate results.

Validation data

Another barrier to uptake of approaches to account for measurement error is the need for validation data. Many analytic approaches to address measurement error rely on validation data, which includes both the error-prone outcome measurement and a gold-standard measurement. These data provide information about the measurement error process, such as the sensitivity and specificity. Ideally these validation data would be available on a (perhaps stratified) random sample of the population included in the study. However internal validation data such as these can be expensive and time consuming to collect. A convenient and less costly alternative is using data that were collected for other purposes but incidentally have both the error-prone and gold-standard measurements (i.e., external data). However, using external data as validation has unique challenges. Specifically, we need to carefully consider whether the information on the measurement error process in the external validation data can be transported to the study population.[61(sec2.2.4)] There is a growing literature on approaches to transport causal effects from one population to another,[64,65] however approaches to transport measurement error parameters from validation data have been only recently explored. Recent work by Edwards et al. and Ackerman et al. explore assumptions and approaches for transporting measurement error parameters when the parameter of interest is the outcome risk.[66,67] Given the convenience and availability of potential external validation data, more work is needed to understand when and how these data can be leveraged for measurement error correction.

Missing data

Missing data plagues nearly all epidemiologic research including pregnancy outcomes research. Unlike hidden measurement error, explicit missing data is observable. Thus, researchers are generally more aware of the negative consequences of inappropriately handling missing data, including bias and loss of efficiency. Despite this, reviews of the epidemiologic and clinical literature show that missing data are often inadequately reported and that complete case analysis, where records with missing data are

excluded from the analysis, remains most frequently implemented.[68–73] Complete case analysis is inefficient because some data are not used in the analysis and it is only unbiased under strong assumptions that are unlikely to hold in realistic settings.[74,75] There are alternatives to complete case analysis with weaker assumptions.

Missing data types

Data may be characterized by the independence, conditionally or marginally, of missingness on observed and missing data.[76,77] Data are "missing completely at random" (MCAR) when missingness is marginally independent of observed and missing data. Data are "missing at random" (MAR) when missingness is independent of missing data conditional on only observed data. That is, missingness cannot depend on data that are unobserved. Finally, if data are neither MCAR nor MAR, then data are "missing not at random" (MNAR). In general, unless data are missing by design, the missing data mechanism is not known and thus whether the data are MAR or MNAR is an untestable assumption. Complete case analysis is unbiased when data are MCAR; complete case analysis may be valid under some settings when data are MAR or MNAR, though such settings may not be common or realistic.[74,75] Alternative approaches for missing data generally require that data are MAR but will also be consistent under the stronger assumption of MCAR.[76,78]

Missingness patterns

Missingness can be categorized as monotone or nonmonotone. In monotone missingness, there is an ordering of the variables in which a variable is only observed if the previous variable was observed.[79] A monotone pattern is most commonly seen when there is a lost to follow-up. Uniform missingness is a special case of monotone missingness in which those variables with missingness are either all observed or missing together. As the name implies, nonmonotone missingness is defined by not having a monotone pattern. When multiple variables have missingness, a nonmonotone pattern is common.

Principled approaches for missing data

Two principled approaches for missing data include multiple imputation (MI) and weighting methods.[78–82] MI and weighting have different assumptions and each approach has advantages and

disadvantages. In MI, the underlying full data is modeled. Conversely, for weighting, the missingness mechanism is modeled, so a distribution of the full data does not need to be specified. For valid estimates, both approaches require that these models (for either the missingness mechanism in weighting or the missing data in MI) are correctly specified. Both approaches require that the missing data are MAR.

Weighting is rather simple to implement when missing data follow a monotone pattern. However, until recently, weighting approaches for a nonmonotone pattern have been very complex.[83,84] In 2018, Sun and Tchetgen Tchetgen published an weighting approach for nonmonotone missing data in the statistical literature.[85] They also published an application in the epidemiologic literature.[81]

Sun and Tchetgen Tchetgen approach

To account for missing data using weighting, the complete cases are weighted by the inverse of their conditional probability of being a complete case. This probability is conditional on the set of covariates required for MAR. Typically, the complete case probability must be estimated. When missingness is uniform, estimation is straightforward and can, for example, be estimated by a logistic regression that includes the fully observed variables. When missingness is nonmonotone, Sun and Tchetgen Tchetgen's approach specifies a logistic model for each missing data pattern except for the complete case pattern. The parameters of these models are estimated by maximizing the joint likelihood. This is the unconstrained maximum likelihood estimator (UMLE). Once these parameters are estimated, the conditional probability for being a complete case is back calculated as the complement of the sum of the probabilities of being in one of the other patterns. Sometimes the UMLE will fail to converge and Sun and Tchetgen Tchetgen propose an alternative constrained Bayesian estimator (CBE).

Unfortunately, there do not appear to be any applications of this approach in the epidemiologic literature (subsequent to 2018), possibly because the available papers are too complex for widespread adoption. weighting for nonmonotone missing data could be a valuable, intuitive, and easy to implement tool for researchers to handle missing data, but the lack of illustrative and easy to follow examples with code are currently barriers to adoption.

Conclusion

Our proposed work will develop, illustrate, and compare the use of innovative tools to produce accurate answers. Researchers are generally aware of the methodological challenges described and yet these issues are rarely addressed in pregnancy outcomes research. This is likely because researchers lack easy to use tools to address these challenges or illustrative examples of implementation of already developed approaches. In this work we will address two important methodological issues that are commonly overlooked. We will illustrate the application of novel tools to applied examples in pregnancy outcomes research in order to promote their uptake by other researchers. It is in part through the development and widespread use of modern methods that we can produce more actionable evidence to improve public health.

CHAPTER 2: STATEMENT OF SPECIFIC AIMS

Annually, there are more than 2.5 million stillbirths, 13.4 million preterm births and 32 million babies born small for gestational age in low- and middle-income countries. These adverse pregnancy outcomes occur disproportionately in low- and middle-income countries and have severe short-term and long-term negative effects. Thus, there is a critical need for effective interventions to prevent these adverse pregnancy outcomes. However, research into such interventions faces methodological challenges that hamper the ability to produce robust and accurate results. These challenges are often ignored, which results in biased or imprecise estimates that misinform public health planning or policy. Informative results are critical for public health programs to make evidence-based decisions, particularly in settings of limited resources.

The overall objective of this dissertation is to address analytic challenges that plague studies of pregnancy outcomes by developing and applying modern approaches for causal inference, and disseminating these tools to researchers through illustrative applications and published code. It is in part through the development and widespread use of modern methods that we can produce more actionable evidence to improve public health. In this proposal, we will tackle two methodological challenges: measurement error and missing data.

In much pregnancy outcomes research, the date of conception of a fetus is unknown and gestational age must be estimated either by the date of the last menstrual period (LMP) as recalled by the pregnant person or by ultrasound. Gestational age measured by ultrasound is more accurate than LMP and it has been shown that using LMP widens the distribution of gestational age at birth compared to ultrasound measurement, resulting in an excess of births classified as preterm. Unfortunately, due to lack of resources, universal measurement by ultrasound is not feasible in many low- and middle-income

countries. Sometimes ultrasound measurement is available in an external population and these data can provide information about the measurement error process. However, to leverage these external data and produce more accurate research results, we must carefully consider differences between populations and the assumptions needed to transport information between them.

Missing data plagues nearly all research. When missingness does not follow strict patterns (i.e., a monotone patterns), most researchers rely on imputation methods that require parametric assumptions for the full data distribution. Recent work has extended semiparametric weighting approaches for such missing data, but there has been limited uptake of this new approach.

Therefore, we specifically aimed to:

1. Examine estimators that leverage external validation data to account for outcome misclassification to estimate marginal risks and causal effects. We used these estimators to account for misclassification in LMP-measured preterm birth in a large, representative population of pregnant people in Lusaka, Zambia captured in an electronic health record, leveraging ultrasound data from external cohorts. We estimated the risk of preterm birth and the effect of maternal HIV infection on preterm birth.
2. Apply a recently developed weighting approach for nonmonotone missingness and compare its finite sample properties with multiple imputation. We focused on the setting of time-fixed inverse probability of treatment weighted marginal structural models and, for illustration, we estimated the effect of anemia on preterm birth.

For this work, we analyzed data from the Zambia Electronic Perinatal Record System (ZEPRS), an electronic medical record used by obstetric clinics in Lusaka, Zambia that includes routine care data on more than 100,000 people; the Zambia Preterm Birth Prevention Study (ZAPPS), an information-rich cohort of 1450 people enrolled during pregnancy in Lusaka; and the Improving Pregnancy Outcomes with Progesterone (IPOP) study, a phase three, randomized, double-blind, placebo-controlled trial of progesterone injection in 800 pregnant people also in Lusaka.

IMPACT: Our work developed, compared, and illustrated the use of innovative analytic tools to produce robust and accurate answers. Such answers are critical for public health programs to make evidence-based decisions on what interventions to pursue to improve population health.

CHAPTER 3: METHODS

Study design

Each aim involves two components: 1) simulation study in which data is generated under a known truth to assess performance of analytic approaches and 2) motivating/applied example in which analytic approaches are implemented in real data sources. The data sources are described in the next section, followed by description of the motivating/applied examples for each aim.

Data sources

We leveraged three existing data sources. ZEPRS was an electronic medical record system used by 25 prenatal clinics and 1 referral hospital in Lusaka, Zambia that includes routine care data on more than 250,000 pregnant people seeking prenatal care between January 2008 and June 2013.[86] ZAPPS was an observational prospective cohort of 1450 people recruited at prenatal care initiation at five of the clinics included in ZEPRS and at the referral hospital between 2015 and 2017.[87,88] IPOP was a phase three, randomized, double-blind, placebo-controlled trial of progesterone injection in 800 pregnant people with confirmed HIV infection and no prior preterm birth enrolled at prenatal care initiation at two of the clinics included in ZEPRS between February 2018 and January 2020.[89,90] Gestational age measured by ultrasound was not available in ZEPRS; rather gestational age was measured either by reported LMP (>78% of singleton pregnancies) or symphysis-fundal height. In ZAPPS and IPOP, participants received high quality ultrasound to measure gestational age at enrollment (prenatal care initiation). In Aim 1, we leveraged the ZAPPS and IPOP cohorts as external validation data in order to account for misclassification of LMP-measured preterm birth in the ZEPRS population. In Aim 2, we implemented weighting for nonmonotone missingness using data from the ZAPPS cohort.

Zambia Electronic Perinatal Record System

ZEPRS was funded by Bill and Melinda Gates Foundation. Field testing of the system began in November 2005 at 3 sites and was expanded in a phased approach.[86] By June 2007, it was implemented across 25 prenatal care clinics, 13 of which were also delivery centers, and 1 referral hospital. The final data extraction from ZEPRS was in 2013. Each patient received a standard identification number at first contact. This identification number followed the patients at different clinics within the system, at delivery, and across pregnancies. At birth, the infant was also assigned a unique identification number that was linked to the mother. Nurses, midwives, and clerical staff entered data from each contact at the point of care in real-time. To ensure data quality and integrity, the system had built-in checks and data quality reports were generated monthly to identify inconsistencies and duplicate entries. The system captured all care at the prenatal clinics and deliveries within the system. The primary strength of ZEPRS is that it captures a population-based sample of pregnancies and the data are representative of pregnant people seeking prenatal care in Lusaka, Zambia.

The Zambian Preterm Birth Prevention Study

ZAPPS, also funded by the Bill and Melinda Gates Foundation, aimed to establish a well-characterized pregnancy cohort to better understand risk factors associated with preterm birth.[87,88] Between August 2015 and September 2017, eligible pregnant people were enrolled at the referral hospital in Lusaka and five nearby high-volume district health clinics. An individual was eligible if they 1) were ≥18 years old, 2) had a viable intrauterine single or twin pregnancy, 3) presented to prenatal care prior to 20 weeks gestation if HIV-uninfected or 24 weeks if HIV-infected, 4) resided within Lusaka with no plans to relocate during follow-up, and 5) provided consent to participate and for the infant to participate. A total of 1784 people were recruited and screened and 1450 were enrolled.

The contact schedule followed the standard of prenatal care in Zambia (4 recommended contacts). At enrollment, a detailed medical history was obtained including information on prior pregnancies and outcomes. At enrollment and each follow-up contact, participants received a physical exam and routine services according to standard of care in Zambia. Participants were compensated for their time and effort

for attendance at study contacts. Participants were encouraged to deliver at the referral hospital and study staff identified participants at admission for labor and delivery. For participants who delivered elsewhere, outcomes were collected in person or by phone.

Improving Pregnancy Outcomes with Progesterone Study

IPOP was a double-masked, placebo controlled, randomized trial of 17-hydroxyprogesterone caproate (17P) to prevent preterm birth (NCT03297216).[89,90] Between February 2018 and January 2020, eligible pregnant people were enrolled at the referral hospital in Lusaka and antenatal clinics of the Kamwala District Health Centre. An individual was eligible if they 1) were ≥18 years old, 2) had a viable intrauterine singleton pregnancy without uterine or fetal anomaly detected on ultrasound, 3) presented to prenatal care prior to 24 weeks, 4) had antibody-confirmed HIV-1 infection, 5) were currently receiving or intended to start antiretroviral therapy in pregnancy, 6) did not have a confirmed prior spontaneous preterm birth, 7) did not have a known allergy or contraindication to 17P, and 8) were able and willing to provide consent and adhere to the weekly study visit schedule. Participants were randomly assigned to one of two arms, weekly intramuscular injection of 17P or placebo starting between 16 and 24 weeks of gestation and continuing until 36 weeks, stillbirth, or delivery. A total of 1042 potentially eligible people were identified and 800 were enrolled and randomized (399 to 17P and 401 to placebo).

After enrollment and randomization, participants had weekly visits for injection. The study also provided routine antenatal care following Zambia guidelines (4 recommended contacts). At enrollment, a detailed medical history was obtained including information on prior pregnancies and outcomes. At enrollment and each antenatal care follow-up contact, participants received a physical exam and routine services according to standard of care in Zambia.

Of the 800 people enrolled, none were lost to follow-up. The primary outcome was a composite of preterm birth or stillbirth. In each arm, 36 individuals had the outcome (9%) for a risk difference of 0.1 percentage points (95% confidence interval -3.9, 4.0). No differences were observed for secondary outcomes including spontaneous preterm delivery, provider-initiated preterm delivery, delivery before 34

weeks, and delivery before 28 weeks. There were also no differences by timing antiretroviral initiation, parity, or gestational age at enrollment.

Gestational age measurement

In ZAPPS and IPOP, at the screening visit (typically the same day as enrollment), participants underwent ultrasound to measure crown-rump length (if <14 weeks by LMP), or head circumference and femur length (if >14 weeks). All measurements were taken twice and averaged. The INTERGROWTH-21st equations were used to estimate gestational age from fetal measurements.[91,92] Pregnancies below the lower threshold of these equations were dated by the Hadlock formula.[93,94] Sonographers were trained using curricula adapted from INTERGROWTH-21st. The primary strength of ZAPPS and IPOP is the capture of accurate gestational age measurement by ultrasound. LMP as recalled by the pregnant person at the first prenatal visit was also documented.

Ultrasound is not part of routine prenatal care in Lusaka, so ZEPRS does not capture gestational age measured by ultrasound. Rather gestational age is mostly commonly measured by reported LMP (>78% of singleton pregnancies).

Motivating/applied examples

Aim 1

In ZEPRS, we aimed to estimate the risk of preterm birth (i.e., the natural course) and the effect, quantified by the risk difference, of maternal HIV infection on the risk of preterm birth. Unfortunately, gestational age measured by ultrasound was not available in ZEPRS; rather gestational age was most commonly measured by reported LMP. Given that LMP-measured gestational age has known error, outcome misclassification is a concern. More recently, two research studies, ZAPPS and IPOP, prospectively enrolled a nonrandom sample of pregnant people at clinics in the ZEPRS system and recorded preterm birth measured by reported-LMP and by ultrasound. We aimed to use these research studies as external validation data to account for outcome misclassification in our analysis of ZEPRS.

Aim 2

In ZAPPS, we aimed to estimate the effect, quantified by the risk difference, of maternal anemia on the risk of spontaneous preterm birth. Some research has suggested an association between maternal anemia, particularly when diagnosed early in pregnancy, and poor pregnancy outcomes.[11,95,96] However, this finding has not been consistently observed.[97,98] Anemia was diagnosed at enrollment if the capillary hemoglobin concentration was <10.5 g/dL (HemoCue Hb 201).[99] Spontaneous preterm birth was defined as delivery occurring after spontaneous labor or membrane rupture prior to 37 weeks of gestation. Additional covariates collected at enrollment to be used in this analysis included gestational age at enrollment, maternal age, maternal HIV serostatus, and previous pregnancy and birth history. Three people experienced a miscarriage and were excluded from the analysis, resulting in 1447 people.

CHAPTER 4: LEVERAGING EXTERNAL VALIDATION DATA: THE CHALLENGE OF TRANSPORTING MEASUREMENT ERROR PARAMETERS

Overview

Approaches to address measurement error frequently rely on validation data to estimate measurement error parameters (e.g., sensitivity and specificity). Acquisition of validation data can be costly, thus secondary use of existing data for validation is attractive. To use these external validation data, however, we may need to address systematic differences between these data and the main study sample. Here, we derive estimators of the risk and the risk difference that leverage external validation data to account for outcome misclassification. If misclassification is differential with respect to covariates that themselves are differentially distributed in the validation and study samples, the misclassification parameters are not immediately transportable. We introduce two ways to account for such covariates in transporting the misclassification parameters: 1) condition on the covariates (i.e., estimate stratified misclassification parameters) or 2) weight the validation sample to match the study sample distribution of the covariates. We provide proofs of identification, describe estimation using parametric models, and assess performance in simulations. We also illustrate implementation to estimate the risk of preterm birth and the effect of maternal HIV infection on preterm birth. Measurement error should not be ignored and it can be addressed using external validation data via transportability methods.

Introduction

Measurement error can produce substantial bias and flawed inference,[47–49,55–58,61] but is often ignored.[100] Approaches to address measurement error typically rely on validation data to estimate measurement error parameters (e.g., sensitivity and specificity).[56,61,101] However, collection of validation data can be costly and time consuming. Thus, secondary use of existing data as validation data is an attractive alternative. As such validation data are necessarily external to the main study sample, we need

to consider how to "transport" measurement error parameters from the validation to the study sample.[61(sec2.2.4),101,102]

The importance of addressing measurement error increases as the use of routinely collected healthcare data for research becomes increasingly common.[103] The primary purpose of these data is individual patient care, so some parameters may be subject to greater measurement error than would be found in a prospective research cohort.[104] The estimation of fetal gestational age in low- and middle-income countries (LMIC) provides an example. In LMIC, gestational age is typically measured by last menstrual period (LMP), which is solicited from the patient during routine prenatal care. Gestational age is a foundational measure that is used to define preterm birth and other important adverse birth outcomes. LMP-derived gestational age is subject to nontrivial measurement error[31–39] that typically results in an overestimate of preterm birth risk.[40,42,51–53] Gestational age derived from early ultrasound is a more accurate measure,[37,41–46] but this technology may not be available outside the research setting. Thus, the outcome of preterm birth may be subject to misclassification in clinical data from LMIC.

In this work, we introduce estimators that leverage external data to account for misclassification of outcomes with particular focus on the challenge of transporting misclassification parameters between populations. We draw on methods for transporting causal effects to address systematic differences between the validation and study samples.[65] We apply the estimators in a motivating example to account for misclassification in preterm birth measured by LMP in a population of women in Lusaka, Zambia captured in an electronic health record, leveraging ultrasound data from external research studies.

Methods

Motivating example

We aimed to estimate the overall risk of preterm birth (i.e., the natural course[105]) and the causal effect of maternal HIV infection on the risk of preterm birth among people seeking prenatal care in Lusaka, Zambia. We use data from the Zambia Electronic Perinatal Record System (ZEPRS), an electronic health record used by 25 clinics and 1 referral hospital in Lusaka with deliveries between January 1, 2008 and June 26, 2013.[86,106] This is our study sample.

In ZEPRS, the outcome of preterm birth was defined by gestational age at birth <37 weeks, as measured by patient-reported LMP. Given that LMP measurement has known error, outcome misclassification is a concern. More recently (2015-2020), two research studies prospectively enrolled a nonrandom sample of pregnant people at clinics where ZEPRS had been deployed.[87,89] These studies assessed gestational age (and thus preterm birth) by both LMP and early ultrasound and are used herein as a validation sample.

Accounting for outcome misclassification

Let Y and A be binary indicators for the true outcome and exposure, respectively. We aimed to estimate, in the study sample, the marginal outcome risk under the natural course and two counterfactual marginal risks: one for the scenario where everyone was exposed and one for the scenario where everyone was unexposed. The natural course is $P(Y = 1|R = 1)$ and the counterfactual risks are $P(Y(a) = 1|R = 1)$, where $Y(a)$ is the potential outcome (i.e., the outcome that would be observed) when $A = a$ and R is an indicator of whether the person was in the study sample, $R = 1$, or in the validation sample, $R = 0$. Let Z be a common cause (or a vector of common causes) of A and Y such that $P(Y(a) = 1|A = a, R = 1) \neq P(Y(a) = 1|R = 1)$. Under conditional exchangeability with positivity and causal consistency,[27,29,107,108] we can identify the counterfactual risks using the g-formula,[27]

$$P(Y(a) = 1|R = 1) = \sum_{z} P(Y = 1|A = a, Z = z, R = 1)P(Z = z|R = 1). \qquad [1]$$

For simplicity, our notation assumes categorical Z.

However, only Y^*, a potentially misclassified version of Y, was measured in our study sample. To account for misclassification, we can replace $P(Y = 1|A = a, Z = z, R = 1)$ in Equation [1] with [109,110]

$$\frac{P(Y^* = 1|A = a, Z = z, R = 1) - \left(1 - Sp_{A,Z,R=1}\right)}{Se_{A,Z,R=1} - \left(1 - Sp_{A,Z,R=1}\right)}, \qquad [2]$$

where $Se_{A,Z,R=1} = P(Y^* = 1|Y = 1, A = a, Z = z, R = 1)$ and $Sp_{A,Z,R=1} = P(Y^* = 0|Y = 0, A = a, Z = z, R = 1)$ are the misclassification parameters sensitivity and specificity, respectively, in the strata of A and Z in the study sample (proof in Appendix 4A).

We can obtain the misclassification parameters from the validation sample (because both Y and Y^* are available), but in our motivating example these data are external and thus conditional on $R = 0$. When R is independent of Y^* conditional on Y, A and Z ($R \coprod Y^* | Y, A, Z$), we can transport the A and Z specific misclassification parameters from the validation sample to the study sample

$$P(Y^* = y^* | Y = y, A = a, Z = z, R = 1) = P(Y^* = y^* | Y = y, A = a, Z = z, R = 0). \quad [3]$$

$R \coprod Y^* | Y, A, Z$ is a conditional transportability condition. **Figure 1A** and **1B** are causal diagrams where this condition holds. We also need positivity, i.e., the validation sample includes individuals across the observed distribution of A and Z in the study sample.[64]

If misclassification is nondifferential with respect to A and Z, ($A, Z \coprod Y^* | Y, R$; Figure 4.1A), then sensitivity and specificity are constant across strata of Z and A (i.e., $P(Y^* = y | Y = y, A = a, Z = z, R = 0) = P(Y^* = y | Y = y, R = 0)$). This condition does not hold in Figure 4.1B, where misclassification is *differential* with respect to A and Z.

Relaxing the transportability condition

When misclassification is differential with respect to non-confounding covariates W and the distribution of W differs between the validation and study samples, then Equation [3] does not hold ($R \not\coprod Y^* | Y, A, Z$; Figure 4.1C). In our motivating example, measurement error in LMP-measured gestational age may be differential by birth history[42] and the distribution of birth history may vary between the validation and study samples (e.g., the proportion nulliparous in each sample may differ). In the presence of these W covariates, the misclassification parameters are transportable if we additionally condition on W,

$$P(Y^* = 1 | Y = 1, A, Z, W, R = 1) = P(Y^* = 1 | Y = 1, A, Z, W, R = 0). \quad [4]$$

This equality holds when $R \coprod Y^* | Y, A, Z, W$. For Figure 4.1C**,** we provide two identification approaches.

In the first approach, we condition on both Z and W so that sensitivity and specificity are also conditional on W (hereafter called *conditioning on* W), $P(Y(a) = 1 | R = 1) =$

$$\sum_{z,w} \frac{P(Y^* = 1|A = a, Z = z, W = w, R = 1) - \left(1 - Sp_{A,Z,W,R=0}\right)}{Se_{A,Z,W,R=0} - \left(1 - Sp_{A,Z,W,R=0}\right)} P(Z = z, W = w|R = 1)$$

where $Se_{A,Z,W,R=r} = P(Y^* = 1|Y = 1, A = a, Z = z, W = w, R = r)$ and $Sp_{A,Z,W,R=r} = P(Y^* = 0|Y = 0, A = a, Z = z, W = w, R = r)$ (proof in Appendix 4B).

In the second approach, we standardize the misclassification parameters to the conditional W distribution in the study sample (i.e., remove the difference in the distribution of W between the two data sources). Specifically,

$$P(Y^* = y|Y = y, A, Z, R = 1)$$

$$= \sum_w P(Y^* = y|Y = y, A = a, Z = z, W = w, R = 1) P(W = w|Y = y, A = a, Z = z, R = 1)$$

$$= \sum_w P(Y^* = y|Y = y, A = a, Z = z, W = w, R = 0) P(W = w|Y = y, A = a, Z = z, R = 1)$$

$$= \sum_w P(Y^* = y|Y = y, A = a, Z = z, W = w, R = 0) P(W = w|A = a, Z = z, R = 1).$$

The first equality is the law of total probability, the 2nd follows equation [4], and the 3rd equality follows from an additional condition that W is conditionally independent of the outcome ($W \coprod Y|A, Z, R$), which we invoke because Y is missing when $R = 1$. These standardized misclassification parameters are then used in Equation [2].

The identification conditions for both approaches, 1) conditioning on W or 2) standardizing the misclassification parameters, hold in Figure 4.1C, but conditions are violated in Figure 4.1D. Regarding the first approach, W is a collider in 4.1D so conditioning on W opens a path between A and Y inducing M-bias (i.e., it violates conditional exchangeability).[111] Therefore, risk differences are not identified, though we can still identify the natural course as conditional exchangeability is not required. Regarding the second approach (standardizing the misclassification parameters), the path between W and Y in 4.1D violates the final condition ($W \coprod Y|A, Z, R$). Therefore, for this approach, none of the parameters of interest are identified.

Estimation

When Z is categorical and there are enough data, we can estimate quantities nonparametrically. Otherwise, we can use parametric models, at the cost of requiring that the models be correctly specified. In Equation [2] we can use a logistic model to estimate $\hat{P}(Y^* = 1|A = a, Z = z, R = 1)$. Alternatively, we can directly estimate $\hat{P}(Y = 1|A = a, Z = z, R = 1)$ by maximizing a modified likelihood of the observed data.[101,112,113] We use this latter approach, which also extends to multinomial outcomes (Appendix 4C). Let $\mu_i = P(Y = 1|A_i, Z_i, R_i = 1)$ where i is an individual level subscript. We specify a logistic model $\mu_i = \text{expit}\big(\beta_0 + \beta_1 A_i + \beta_z h(Z_i)\big)$ where $\text{expit}(\cdot) = 1/\big[1 + \exp\big(-1(\cdot)\big)\big]$ and $h(\cdot)$ is an arbitrary flexible function that may include interactions with other variables. We estimate $\hat{\beta}$ by maximizing the modified likelihood in the study sample

$$L(\beta) = \prod_{i=1}^{n} \left\{ \left[\widehat{Se}_i \times \mu_i + \left(1 - \widehat{Sp}_i\right) \times (1 - \mu_i)\right]^{R_i Y_i^*} \left[\left(1 - \widehat{Se}_i\right)\mu_i + \widehat{Sp}_i(1 - \mu_i)\right]^{R_i\left(1 - Y_i^*\right)} \right\}$$

where $\widehat{Se}$ and $\widehat{Sp}$ are the sensitivity and specificity, respectively (transported from the validation sample), and n is the total number of individuals (combined study and validation samples). To estimate the natural course risk, we take the mean of $\hat{\mu}_i$, $\frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} R_i \hat{\mu}_i$. To estimate the counterfactual risks for each treatment level, $\hat{P}(Y(a) = 1) = \frac{1}{\sum_{i=1}^{n} R_i} \sum_{i=1}^{n} R_i \hat{\mu}(a)_i$ where $\hat{\mu}(a)_i$ is an individual's estimated outcome risk when $A = a$ (i.e., g-computation).[114]

If misclassification is differential with respect to A and continuous or high-dimensional Z, we can use a parametric misclassification model that is conditional on A and Z.[101] We specify a logistic model for misclassification $P(Y^* = 1|Y_i, A_i, Z_i, R_i = 0) = \text{expit}(\delta_0 + \delta_1 Y_i + \delta_2 A_i + \delta_3 A_i Y_i + \delta_z h(Z_i))$ in the validation sample. Subsequently, we transport the fitted $\hat{\delta}$ to the study sample. For each individual in the study sample, $\widehat{Se}_i = \text{expit}(\hat{\delta}_0 + \hat{\delta}_1 \times 1 + \hat{\delta}_2 A_i + \hat{\delta}_3 A_i \times 1 + \hat{\delta}_z h(Z_i))$ and $1 - \widehat{Sp}_i = \text{expit}(\hat{\delta}_0 + \hat{\delta}_1 \times 0 + \hat{\delta}_2 A_i + \hat{\delta}_3 A_i \times 0 + \hat{\delta}_z h(Z_i))$, in which Y is set to 1 and 0, respectively.

If misclassification is also differential with respect to W as in Figure 4.1C, one approach is to condition on W. Now, $\mu_i = P(Y = 1|A_i, Z_i, W_i, R_i = 1) = \text{expit}(\beta_0 + \beta_1 A_i + \beta_z h(Z_i) + \beta_w h(W_i))$; the misclassification model is $P(Y^* = 1|Y_i, A_i, Z_i, W_i, R_i = 0) = \text{expit}(\delta_0 + \delta_1 Y_i + \delta_2 A_i + \delta_3 A_i Y_i + \delta_z h(Z_i) + \delta_w h(W_i))$. A second approach is to standardize the misclassification parameters by W. If W is continuous or high-dimensional, estimation is challenging; however, there is an equivalent weighted estimator and we take this approach going forward (proof Appendix 4D).[66] The weight is the stabilized odds of selection,

$$\pi = \frac{P(R = 1|W = w, A = a, Z = z)}{P(R = 0|W = w, A = a, Z = z)} \frac{P(R = 0|A = a, Z = z)}{P(R = 1|A = a, Z = z)}.$$

To estimate the weight, we specify two logistic selection models $P(R_i = 1|A_i, Z_i, W_i) = \text{expit}(\nu_0 + \nu_1 A_i + \nu_z h(Z_i) + \nu_w h(w_i))$ and $P(R_i = 1|A_i, Z_i) = \text{expit}(\phi_0 + \phi_1 A_i + \phi_z h(Z_i))$. Now, the misclassification model is $P(Y^* = 1|Y_i, A_i, Z_i, R_i = 0) = \text{expit}(\delta_0 + \delta_1 Y + \delta_2 A_i + \delta_3 A_i Y_i + \delta_z h(Z_i))$, fit in the validation sample weighted by π.

Estimation of the variance of our parameters of interest must account for uncertainty in the misclassification parameters estimated in the validation sample. One option is nonparametric bootstrap in which both data sets are independently resampled. Alternatively, in this work, we use M-estimation and the empirical sandwich variance estimator,[115,116] which is more computationally efficient than bootstrap (Appendix 4E).

Simulations

We conducted illustrative simulations. We generated data loosely based on our motivating example for 5000 cohorts under each of the four scenarios depicted in Figure 4.1 (Appendix 4F, code available at https://github.com/rachael-k-ross/MeasurementError-ExternalValidation). For each scenario, we conducted six analyses using the estimators described above to estimate the natural course outcome risk and the marginal causal risk difference: 0) using the true outcome Y (although impossible in practice, this provides a benchmark); 1) using the misclassified outcome Y^*; 2) accounting for nondifferential misclassification in Y^*; 3) accounting for differential misclassification with respect to A and Z; and

accounting for differential misclassification with respect to A, Z, and W by 4) conditioning on W or 5) weighting the misclassification parameters. To describe performance, we estimated bias, empirical standard error (ESE), average estimated standard error, and 95% confidence interval (CI) coverage.[117]

Applied example

As previously described, we aimed to estimate the risk of preterm birth and the effect (marginal risk difference) of maternal HIV infection on preterm birth in the study sample.

Study sample

We restricted the ZEPRS study sample to individuals aged 18 to 40 at the first prenatal care visit with a singleton delivery, with gestational age measured by LMP, who initiated prenatal care <24 weeks gestation, and with gestational age at delivery ≥16 weeks (to overlap with validation sample and exclude spontaneous abortions). We also excluded pregnancies with questionable documentation of gestational age (i.e., first visit occurred after delivery, gestational age at first visit ≤2 weeks, gestational age at birth ≥46 weeks) and pregnancies missing data on HIV status. The final analytic cohort included 98,805 pregnancies. HIV status was captured by patient report or by rapid testing at or prior to the first visit. Maternal age was *a priori* deemed to be a confounder.

Validation sample

We used data from the Zambian Preterm Birth Prevention Study (ZAPPS) and the Improving Pregnancy Outcomes with Progesterone (IPOP) trial for validation. ZAPPS was an observational prospective cohort of 1450 people enrolled between August 2015 and September 2017 at prenatal care initiation at five clinics that had been included in ZEPRS.[87] IPOP was a phase three, randomized, double-blind, placebo-controlled trial of weekly progesterone injection in 800 people with confirmed HIV infection and no prior spontaneous preterm birth enrolled between February 2018 and January 2020 at prenatal care initiation at two clinics that had been included in ZEPRS.[89] Progesterone had no effect on preterm birth in this trial.[90] In both studies, gestational age was measured by ultrasound using the INTERGROWTH-21st dating equations.[91,92] LMP was also recorded. Here, we treat ultrasound-measured preterm birth as the gold standard. HIV status was obtained by rapid test at enrollment. We excluded non-

singleton pregnancies; spontaneous abortions; and pregnancies with missing data on delivery date, HIV status, or maternal age. The final validation sample included 1778 pregnancies (1026 ZAPPS and 752 IPOP).

Analysis

Appendix 4G is a causal diagram. We conducted five analyses that parallel those in the simulation: (1) naïve analysis using LMP-measured preterm birth; (2) accounting for nondifferential outcome misclassification; (3) accounting for differential outcome misclassification with respect to HIV status and maternal age; and accounting for differential outcome misclassification with respect to HIV status, maternal age, nulliparity, and prior preterm birth by (4) conditioning on these variables or (5) weighting the misclassification model by nulliparity and prior preterm birth. Henceforth, we refer to nulliparity and prior preterm birth collectively as birth history. Maternal age was modeled using restricted quadratic splines (4 knots at 5th, 35th, 65th, and 95th percentiles);[118] other variables were binary. Models included all two-way interaction terms (interactions with age included linear term only).

Results

Simulations

The true natural course risk and risk difference were 23.3% and 5.0 percentage points, respectively. Table 4.1 and Figure 4.2 show simulation results.

In scenario A (nondifferential misclassification), the naïve analysis (1) was biased upward (+9.2 percentage points) for the natural course risk and biased downward for the risk difference (-1.2). The estimator accounting for nondifferential misclassification (2) had negligible bias. The estimators accounting for differential misclassification (3-5) also had negligible bias, though these estimators were less efficient, particularly for the risk difference (ESE ~4 times larger).

In scenario B (differential misclassification by exposure and confounder), both the naïve analysis (1) and the estimator accounting for nondifferential misclassification (2) were biased. Bias in the risk difference was larger for the estimator accounting for nondifferential misclassification than for the naïve

analysis (+6.6 vs. +3.7). Estimators accounting for differential misclassification (3-5) had negligible bias. The estimator that weighted the misclassification parameters by W (5) was the least efficient.

In scenario C (differential misclassification by exposure, confounder, and W), only the two approaches that accounted for W (4,5) had negligible bias for both parameters. Conditioning on W (4) was more efficient than weighting the misclassification parameters (5).

In scenario D (differential misclassification by exposure, confounder, and W, and W was a collider), we expected both estimators accounting for W (4,5) to be biased for the risk difference, however, they had negligible bias. We also expected the weighted estimator (5) to be biased for the natural course risk, however, it had negligible bias. Recall that when W is a collider, conditioning on W induces M-bias. It has been shown that meaningful M-bias requires strong relationships along the "M" path.[111] Therefore, we examined bias under an altered data generation mechanism with strong effects (odds ratio 4) for each arrow on the "M" path (Appendix 4H) and, as expected, the estimator conditioning on W was biased for the risk difference and there was negligible bias for the natural course risk. However, unexpectedly, the weighted estimator remained unbiased. To further investigate the potential bias of this weighted estimator, we conducted a simplified simulation without A or Z focused on estimating the outcome risk in the presence of a W covariate (Appendix 4I). In this simplified setting, we observed bias in this estimator when there was a direct path from W to Y.

For all estimators the average estimated standard error was close to the ESE (Table 4.1). Confidence intervals had nominal coverage for unbiased estimators. These results indicate that the standard errors estimated by the empirical sandwich variance estimator appropriately captured random error.

In Table 4.1 we used stabilized weights for the weighted estimator (5). Using unstabilized weights was more likely to fail and was notably less efficient than using stabilized weights (Appendix 4J). Additionally, the estimated standard errors did not appropriately capture the inflated ESE for the unstabilized weighted estimator.

Applied example

In the study sample, 23.0% were HIV positive (22759/98805); 56.8% of the validation sample were HIV positive (1010/1778). Overall, the study sample was younger than the validation sample (e.g., proportion <20 years: 13.5% vs. 4.8%) (Table 4.2). In both samples people with HIV were, on average, older than people living without HIV. There were more nulliparous people (40.5% vs. 25.8%) and fewer with a prior preterm birth (2.4% vs. 16.9%) in the study sample than the validation sample.

Marginally, the sensitivity was 0.90 and specificity was 0.84 in the validation sample. Figure 4.3 presents the results of different misclassification modeling strategies that reflect the three estimators that accommodate differential misclassification. When misclassification was modeled as a function of HIV status and maternal age (A and B), sensitivity was differential by HIV status only at younger ages; specificity was differential by HIV status, but varied little by age. When misclassification was also modeled as a function of birth history (C and D), parous individuals were the only group with notable variation of sensitivity by age and HIV status; there was variation in specificity by birth history and age only among people with HIV. Panels E and F show misclassification modeled as a function of HIV status and age, weighted so that the validation sample had the same birth history distribution as the study sample. Weighting had little impact except on sensitivity among HIV negative individuals.

Table 4.3 presents the estimates (in percentage points) of the natural course risk, the two counterfactual risks (if everyone had HIV and if no one had it), and the risk difference. In the naïve analysis, the natural course was 38.9% (95%CI 38.3, 38.9) which is expected to be biased upward. Accounting for nondifferential misclassification reduced the natural course to 30.3% (95%CI 27.8, 32.8). Estimates of the natural course varied between 32.4% and 34.0% when accounting for differential misclassification. The naïve risk difference estimate was 6.0% (95%CI 5.3, 6.8). The estimate was 8.2% (95%CI 7.0, 9.4) when accounting for nondifferential misclassification. The estimate was 2.3% (95%CI -3.9, 8.5) when accounting for differential misclassification by HIV status and age and the point estimates changed little when also accounting for birth history.

The confidence intervals from the naïve analysis were narrow. The confidence interval width increased when accounting for the misclassification, particularly under the assumption of differential misclassification. When also accounting for birth history, the confidence interval width was larger when conditioning than when weighting; contrary to the simulation. Conditioning on birth history required fitting a misclassification model that included a large number of variables in the validation sample resulting in greater uncertainty given the validation sample size.

Discussion

In epidemiology there is increasing interest in methods to transport causal effects from one population to another.[119,120] In this paper, we draw on those methods in the context of leveraging external validation data to address measurement error. Here, we need to transport misclassification parameters like sensitivity and specificity from an external validation sample to our study sample. Our estimators rely on an exchangeability condition that we call the conditional transportability condition (among others) and we illustrate that causal diagrams can be used to assess conditions for transportability. For estimation, our work incorporates common transportability methods, such as inverse odds weighting, into parametric approaches for measurement error.[65,101]

The measurement error literature focuses more on mismeasured exposures and covariates than it does on outcomes.[54,56] Our simulations highlight that outcome misclassification can produce substantial bias in risks and causal effects. Combined with the high precision of the naïve analysis, ignoring measurement error results in a high degree of confidence in a biased estimator. Our simulations also show that assuming misclassification is nondifferential when it is actually differential can produce more bias than ignoring the misclassification.

Combining multiple data sources like validation sample and a study sample yields a fusion design.[116] When fusing data sources, we should carefully consider assumptions for transporting information between them. Here, we transported misclassification parameters from the validation sample to the study sample.[61(sec2.2.4)] Discussions of differential measurement error in the literature are not often linked to transportability. Additionally, the literature predominantly focuses on whether measurement

error is differential with respect to the exposure (in the context of outcome measurement) and confounders.[56] Whether measurement error is differential with respect to other covariates has largely been ignored (Edwards et al. is a recent exception).[67] To transport the misclassification parameters we need to consider whether measurement error is differential with respect to any non-confounding covariates (W) with differing distributions between the validation and study samples.

We introduced two estimators to account for these W covariates: 1) conditioning on W and 2) weighting the misclassification model by W (using inverse-odds weights commonly seen in the transportability literature[65,121]) so that the validation sample distribution of W is the same as the distribution in the study sample (see Ackerman et al. for single arm trial setting[66]). Unfortunately, when W causes the outcome or is a collider, assumptions required for these approaches are violated so some of our parameters of interest may not be identified. However, in simulations, there was negligible bias under our original data generation based on our motivating example. To observe bias, we had to generate data with certain strong relationships that may be unrealistic.[111] For the weighted estimator, we were only able to observe bias under a simplified version of our simulation. Given that these two proposed approaches rely on different assumptions, similar results may indicate minimal bias. In our application, LMP measurement error was potentially differential by certain elements of the birth history (parity and prior preterm birth) and the distribution of these covariates differed between the study and validation samples. However, birth history may be a collider (Appendix 4G). Ultimately, the results from conditioning on birth history or weighting the misclassification model were similar; they were also similar to the analysis that only accounted for differential misclassification by HIV status and maternal age, therefore accounting for birth history may not be needed.

Our work highlights that to transport we generally require rich validation data sources that capture exposure, confounders, and other covariates related to measurement error. Such data may be rare and costly to collect. When addressing outcome misclassification in electronic health records or other large patient-care datasets, leveraging previously-collected data from prospective research studies that enrolled people from the same health care delivery network may be convenient and low cost. However, if

there are strict eligibility criteria or barriers to participation, then the transportability assumptions may be harder to meet (e.g., there may be positivity violations that require restriction of the study sample).

Our work has limitations. The results of our applied example did not match our expectations. We hypothesized a natural course risk <20% and a risk difference >5 percentage points based on prior analysis of the ZAPPS data.[122] Our results could be biased because we did not meet the transportability assumption (i.e., there were unadjusted W covariates such a pre-pregnancy weight, education, or gestational age at first prenatal care visit). However, even multiple unadjusted W covariates may not be strong enough to notably change estimates after accounting for differential misclassification by HIV status and maternal age. Additionally, our validation sample could have systematic error (e.g., ultrasound technicians were not blinded to the LMP-measured gestational age). Finally, there is likely measurement error in other variables, such as HIV status or prior preterm birth.

Conclusions

Measurement error poses a threat to descriptive and causal parameters. Qualitative conjectures about the direction of potential bias are often based on unrealistic assumptions and do not produce more accurate results. To appropriately address misclassification, we generally need rich and large validation data to meet identification conditions, which is a barrier to widespread adoption of measurement error corrections. We proposed estimators that leverage potentially low-cost and readily available external data.

Figures

Figure 4.1. Causal diagrams for simulation scenarios. ϵ_y is the measurement error. Arrows from R into other nodes signify that the distributions of A, Z, Y, and W differ between the external validation data and study sample indicated by R.

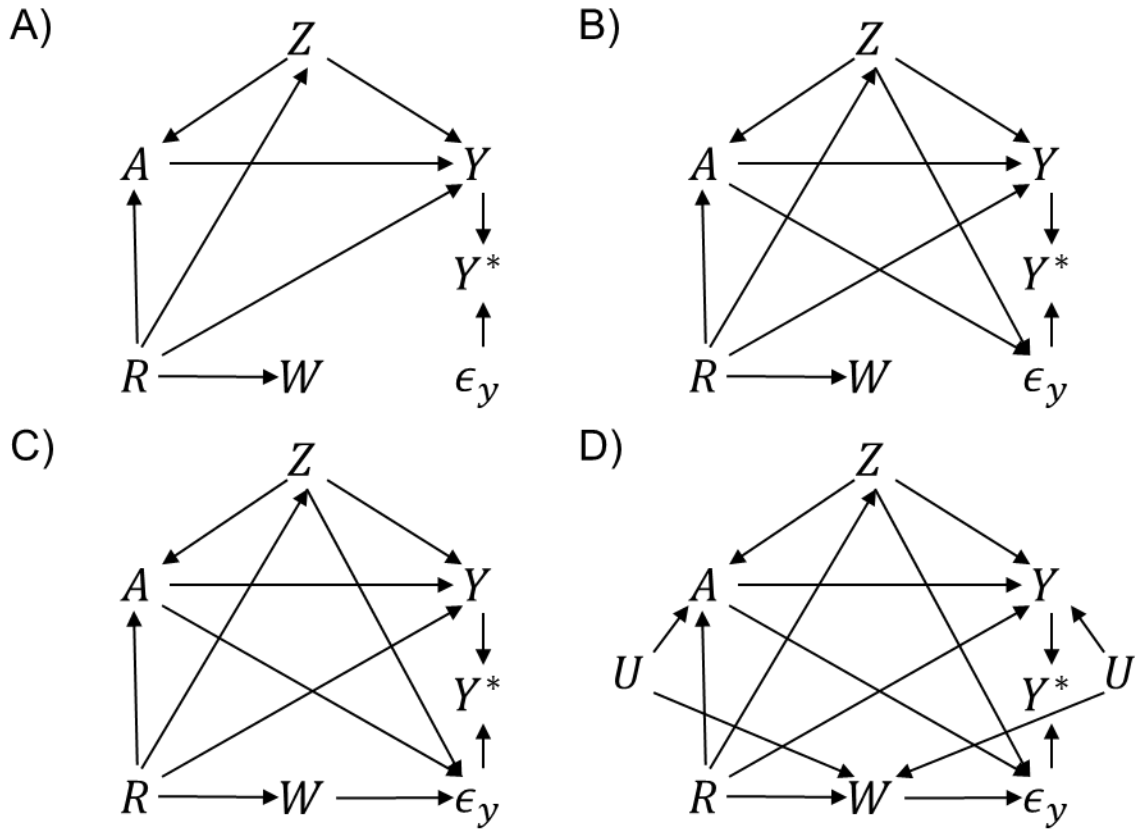


Figure 4.2. Boxplots of risk under the natural course (panel A) and risk difference (B) estimates for simulation under original data generation parameters. Panel A) true risk was 23.2 percentage points; panel B) true risk difference was 5 percentage points. Analytic approaches: 0) true outcome; 1) naïve analysis; 2) accounting for nondifferential error; 3) accounting for differential error by exposure and confounder; and accounting for differential error by exposure, confounder and covariate W by 4) conditioning on W and 5) weighting misclassification parameters by W. Horizontal black line marks true risk and risk difference; black dot marks mean of the estimates; small gray dots are a 10% random sample of estimates.

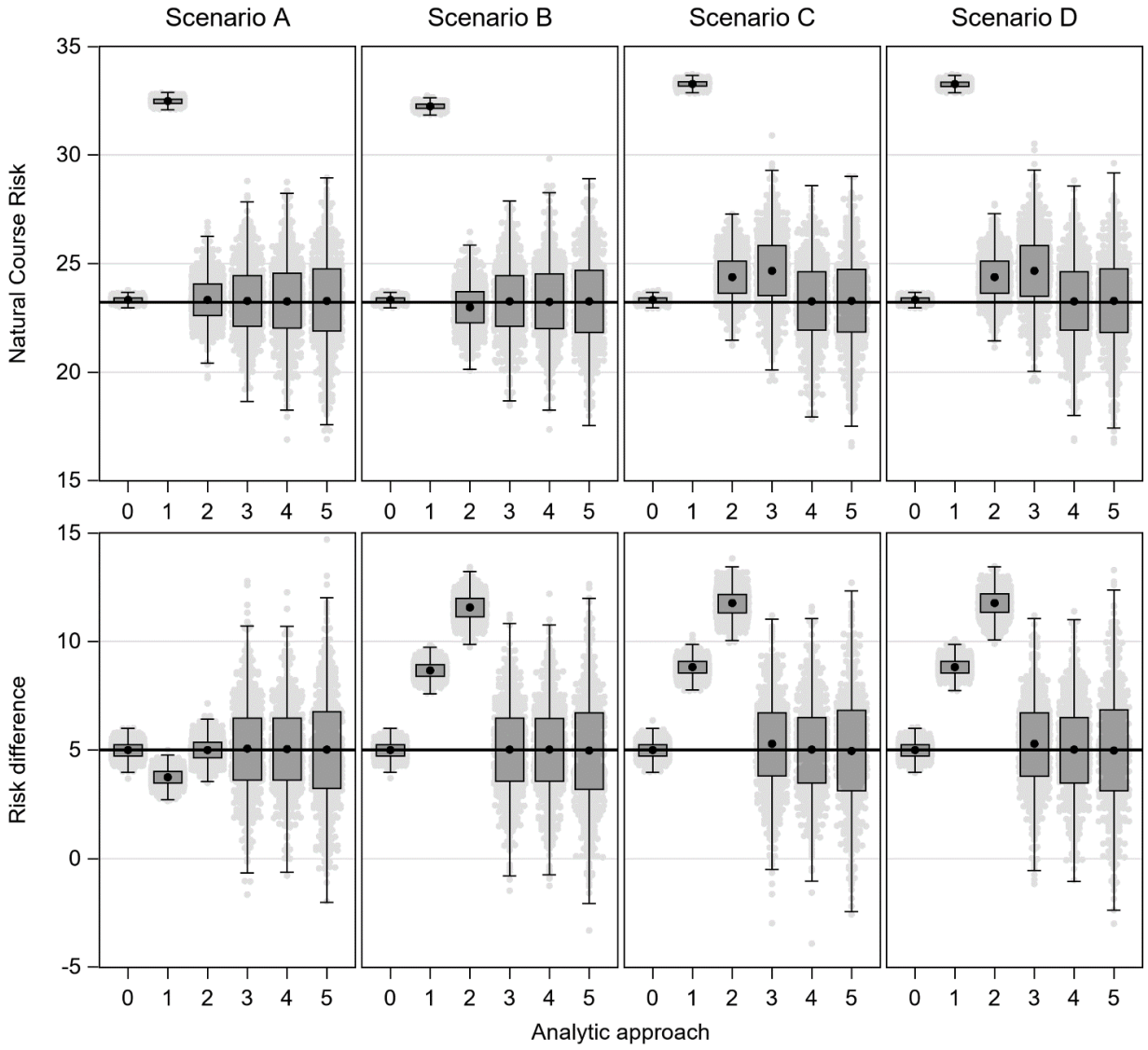
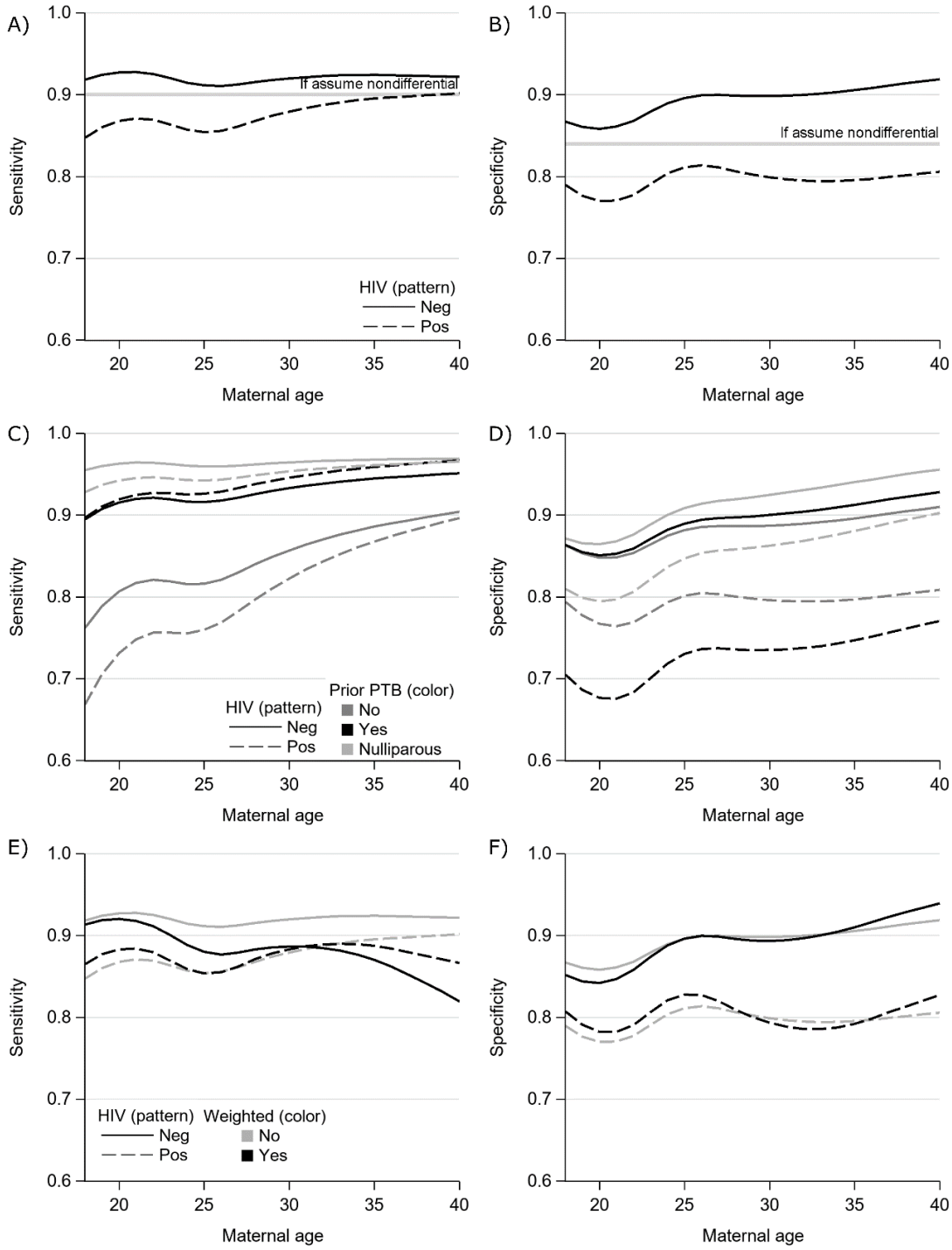


Figure 4.3. Estimated sensitivity (panels A,C,E) and specificity (B,D,F) transported from the validation data to the study sample when misclassification parameters are differential with respect to HIV status and maternal age (A,B), differential with respect to HIV status, maternal age, and birth history (C,D), and differential with respect to HIV status and maternal age weighted by birth history (E,F). The gray lines in panels E and F are the same as the black lines in panels A and B. Abbreviations: Neg, negative; Pos, positive; PTB, preterm birth.



Tables

Table 4.1. Simulation results for the risk under the natural course and the risk difference (in percentage points) under the original data generation (n=5000).

		Natural course					Risk difference				
Scenario[1]	Approach[2]	Mean	Bias	ESE	Avg SE	Coverage[3]	Mean	Bias	ESE	Avg SE	Coverage[3]
All	True outcome (0)	23.3	0.0	0.1	0.1	95.3	5.0	0.0	0.4	0.4	94.4
A	Naïve analysis (1)	32.5	9.2	0.1	0.1	0.0	3.8	-1.2	0.4	0.4	12.5
	Accounting for error that is										
	Nondifferential (2)	23.3	0.0	1.1	1.1	95.5	5.0	0.0	0.5	0.6	94.9
	Differential by A & Z (3)	23.3	0.0	1.7	1.7	95.6	5.1	0.1	2.1	2.2	95.4
	Differential by A, Z, & W										
	Conditioning (4)	23.3	-0.1	1.9	1.9	95.4	5.1	0.1	2.1	2.2	95.4
	Weighted Se/Sp (5)	23.3	0.0	2.1	2.1	95.6	5.0	0.0	2.6	2.6	95.0
B	Naïve analysis (1)	32.2	8.9	0.1	0.1	00.0	8.7	3.7	0.4	0.4	0.0
	Accounting for error that is										
	Nondifferential (2)	23.0	-0.3	1.1	1.1	93.9	11.6	6.6	0.6	0.6	0.0
	Differential by A & Z (3)	23.3	-0.1	1.7	1.7	95.2	5.0	0.0	2.2	2.2	95.5
	Differential by A, Z, & W										
	Conditioning (4)	23.2	-0.1	1.9	1.9	95.4	5.0	0.0	2.2	2.2	95.6
	Weighted Se/Sp (5)	23.3	-0.1	2.1	2.1	95.8	5.0	0.0	2.7	2.7	95.4
C	Naïve analysis (1)	33.3	10.0	0.1	0.1	00.0	8.8	3.8	0.4	0.4	0.0
	Accounting for error that is										
	Nondifferential (2)	24.4	1.0	1.1	1.1	84.8	11.8	6.8	0.6	0.6	0.0
	Differential by A & Z (3)	24.7	1.3	1.7	1.7	86.9	5.3	0.3	2.2	2.2	95.2
	Differential by A, Z, & W										
	Conditioning (4)	23.2	-0.1	1.9	1.9	95.7	5.0	0.0	2.3	2.3	95.4
	Weighted Se/Sp (5)	23.3	-0.1	2.2	2.2	95.4	5.0	0.0	2.8	2.8	94.8

D	Naïve analysis (1)	33.3	10.0	0.1	0.1	0.000	8.8	3.8	0.4	0.4	0.000
	Accounting for error that is										
	Nondifferential (2)	24.4	1.0	1.1	1.1	0.849	11.8	6.8	0.6	0.6	0.000
	Differential by A & Z (3)	24.7	1.3	1.7	1.7	0.869	5.3	0.3	2.2	2.2	0.952
	Differential by A, Z, & W										
	Conditioning (4)	23.2	-0.1	1.9	1.9	0.957	5.0	0.0	2.3	2.3	0.954
	Weighted Se/Sp (5)	23.3	-0.1	2.2	2.2	0.953	5.0	0.0	2.8	2.8	0.948

Abbreviations: ESE, empirical standard error; Avg SE, average estimated standard error; Se, sensitivity; Sp, specificity

[1]Scenarios correspond to Figure 1 Panels A, B, C, and D

[2]Numbers in parentheses correspond to numbered approaches in Figure 4.2

[3]95% confidence interval coverage (%)

Table 4.2. Characteristics of study sample and validation data in applied example, overall and by HIV status.

	Study sample			Validation data		
	HIV status			HIV status		
Characteristic	Positive (N=22759)	Negative (N=76046)	Overall (N=98805)	Positive (N=1010)	Negative (N=768)	Overall (N=1778)
Maternal age						
Median (IQR)	27 (23, 31)	24 (21, 28)	25 (21, 29)	29 (25, 33)	26.5 (23, 31)	28 (24, 33)
<20 years	1360 (6.0)	12016 (15.8)	13376 (13.5)	28 (2.8)	58 (7.6)	86 (4.8)
20-34 years	19004 (83.5)	59592 (78.4)	78596 (79.5)	784 (77.6)	613 (79.8)	1397 (78.6)
≥35 years	2395 (10.5)	4438 (5.8)	6833 (6.9)	198 (19.6)	97 (12.6)	295 (16.6)
Prior preterm birth						
Yes	799 (3.5)	1603 (2.1)	2402 (2.4)	97 (9.6)	203 (26.4)	300 (16.9)
No, parous	15515 (68.2)	40872 (53.7)	56387 (57.1)	726 (71.9)	294 (38.3)	1020 (57.4)
No, nulliparous	6445 (28.3)	33571 (44.1)	40016 (40.5)	187 (18.5)	271 (35.3)	458 (25.8)

Table 4.3. Estimated risks and risk differences, in percentage points, from the applied example.

	Risks (95% CI)			
Analysis	Natural Course	HIV +	HIV -	Difference (95% CI)
Naïve analysis	38.6 (38.3, 38.9)	43.3 (42.6, 44.0)	37.3 (36.9, 37.6)	6.0 (5.3, 6.8)
Accounting for error that is				
Nondifferential	30.3 (27.8, 32.8)	36.7 (33.9, 39.5)	28.5 (26.0, 30.9)	8.2 (7.0, 9.4)
Differential by HIV, age	32.4 (29.6, 35.1)	34.3 (28.8, 39.8)	32.1 (29.0, 35.2)	2.3 (-3.9, 8.5)
Differential by HIV, age, birth history[1]				
Conditioning	34.0 (29.6, 38.5)	36.1 (29.7, 42.4)	33.8 (28.4, 39.2)	2.3 (-5.7, 10.3)
Weighted sensitivity/specificity	32.8 (29.2, 36.4)	34.8 (29.4, 40.3)	32.6 (28.3, 37.0)	2.2 (-4.7, 9.1)

[1]Nulliparity, prior preterm birth

CHAPTER 5: ACCOUNTING FOR NONMONOTONE MISSING DATA USING INVERSE PROBABILITY WEIGHTING

Overview

Inverse probability weighting can be used to correct for missing data. New estimators for the weights in the nonmonotone setting were introduced in 2018. These estimators are the unconstrained maximum likelihood estimator (UMLE) and the constrained Bayesian estimator (CBE), introduced as an alternative if UMLE fails to converge. In this work we describe and illustrate these estimators, and examine their performance in simulation and in an applied example estimating the effect of anemia on spontaneous preterm birth in the Zambia Preterm Birth Prevention Study. We compare performance with multiple imputation (MI) and focus on the setting of an observational study where inverse probability of treatment weights are used to address confounding. In simulation, weighting was less statistically efficient at the smallest sample size and lowest exposure prevalence examined (n=1500, 15% exposure prevalence) but in other scenarios statistical performance of weighting and MI was similar. Weighting had improved computational efficiency taking, on average, 0.4 and 0.05 times the time for MI in R and SAS, respectively. UMLE was easy to implement in commonly used software and convergence failure occurred just twice in >200,000 simulated cohorts making implementation of CBE unnecessary. In conclusion, weighting is a viable alternative to MI for nonmonotone missingness. It may be preferred with large sample sizes, when using resampling algorithms for variance, or when researchers having greater confidence in correctly modeling the missingness mechanism than the missing data values. Weighting for missing data may be more intuitive for researchers already familiar with weighting approaches for other biases.

Introduction

Missing data plague research. Reviews of the epidemiologic and clinical literature show that missing data are often inadequately reported and that complete case analysis, where records with missing data are excluded, remains the most frequently implemented approach to handle missing data.[68–73] Complete case analyses are statistically inefficient and are valid only under strong assumptions.[74,75] Weighting is an alternative approach to handle missing data that is valid under weaker assumptions.[78,79,82]

It is generally straightforward to estimate weights to account for missing data when missing data follow a uniform pattern (i.e., for each individual, the variables with missingness are either all observed or all missing) or a monotone pattern (i.e., there is an ordering in which a variable is observed only if the previous variable is observed, such as missing data after lost to follow-up) (see illustration in Appendix 5A).[79] However, until recently, weighting approaches for nonmonotone missing data (i.e., when missingness is neither uniform nor monotone) have been challenging to implement.[83,84] In 2018, Sun and Tchetgen Tchetgen published two estimators for weights in the setting of nonmonotone missingness.[81,85] Unlike prior approaches, their estimators can be readily implemented in commonly used software.

In this paper, we describe and illustrate the estimators from Sun and Tchetgen Tchetgen, and examine their performance in simulation and an applied example estimating the effect of anemia on spontaneous preterm birth in the Zambia Preterm Birth Prevention Study (ZAPPS).[87,88] We compare performance with multiple imputation (MI), a commonly used alternative. We specifically examine the setting of an observational study where inverse probability of treatment weights are used to address confounding. In section 2, we introduce our motivating application. In section 3, we detail our parameter of interest, a sufficient set of identification assumptions, and weighted estimators. In section 4, we describe the weighted estimators from Sun and Tchetgen Tchetgen using a simple example to aid understanding. In section 5, we present results from a limited simulation study to compare performance with MI in finite samples. Section 6 presents results from the motivating application. Finally, in section 7, we discuss the findings and consider the choice between weighting and MI.

Motivating application

Our objective was to estimate the effect of maternal anemia on the risk of spontaneous preterm birth among people seeking prenatal care in Lusaka, Zambia. Some research has suggested an association between maternal anemia, particularly when diagnosed early in pregnancy, and poor pregnancy outcomes.[11,95,96] However, this finding has not been consistently observed.[97,98] To estimate this effect, we used data from ZAPPS,[87,88] an observational prospective cohort of 1450 people recruited at prenatal care initiation in Lusaka, Zambia between 2015 and 2017. A person was eligible if she was ≥18 years old, had a viable intrauterine single or twin pregnancy, presented to prenatal care prior to 20 weeks of gestation if HIV-seropositive or 24 weeks if HIV-seronegative, and resided within Lusaka with no plans to relocate during follow-up. Anemia was diagnosed at enrollment if the capillary hemoglobin concentration was <10.5 g/dL (HemoCue Hb 201).[99] Spontaneous preterm birth was defined as delivery occurring after spontaneous labor or membrane rupture prior to 37 weeks of gestation. Additional covariates collected at enrollment and used in this analysis included gestational age at enrollment, maternal age, maternal HIV serostatus, and previous pregnancy and birth history. Three people experienced a miscarriage and were excluded from the analysis, resulting in 1447 people.

Table 5.1 shows cohort characteristics and occurrence of the outcome, overall and stratified by anemia diagnosis. ZAPPS is typical of many prospective cohorts. Despite rigorous study procedures and active efforts to maximize study retention, some data are missing. In particular, our exposure and outcome have notable missingness; 425 people (29%) did not have a hemoglobin measurement and 239 (17%) were lost to follow-up. Ignoring missing data, 13.5% of people were anemic and 9.9% had a spontaneous preterm birth. The risk of spontaneous preterm birth was higher among anemic people (12.4%) compared to people without anemia (9.5%) or people with missing anemia status (9.7%). There was also missingness in some covariates: maternal age (n=41, 3%), maternal HIV serostatus (n=3, <1%), and prior stillbirth (n=85, 6%). Table 5.2 shows the 16 missing data patterns for the 5 variables with missingness. Just over half (781, 54%) of the cohort were complete cases. Among the 666 people with some missing data, 4 patterns accounted for 88%. There were 9 patterns with <1% of people.

Parameter, identification, and weighted estimators

Parameter

Our parameter is the sample average causal effect of a time-fixed binary exposure on the outcome risk, quantified by the risk difference, $\theta = E(Y^{x=1}) - E(Y^{x=0})$, where Y^x is the potential outcome when exposure X is set to x. This parameter requires identification of two sample average risks, $E(Y^x)$, one under exposure ($x = 1$) and one under no exposure ($x = 0$). We focus on identification of the risk in the observational setting in which there are common causes of the exposure and treatment, Z, that produce confounding bias such that the risk is not identified by the crude conditional risk, i.e., $E[Y^x] \neq E[Y|X = x]$, even in the absence of missing data.

No missing data

We can point-identify the risk under the assumptions of conditional exchangeability with positivity, causal consistency, and no measurement error. Conditional exchangeability means that, conditional on a set of measured confounders, the potential outcomes are independent of the observed exposure, $Y^x \coprod X|Z$, such that $E[Y^x|Z] = E[Y^x|X = x, Z]$.[29,123] Positivity means that every person has a non-zero probability of having each level of exposure across the distribution of Z.[29,107,123] Causal consistency means that the potential outcome Y^x is the observed outcome Y for people with observed exposure x.[29,108] With these conditions, $E(Y^x)$ is identified by a weighted risk where the weight is the inverse of the confounder-conditional probability of exposure (i.e., inverse probability of treatment weight, hereinafter *treatment weight*), formally

$$E[Y^x] = E\left[\frac{YI(X = x)}{\Pr(X = x|Z = z)}\right],$$

where $I(a)$ is an indicator that takes the value 1 when a is true and 0 otherwise (proof in Appendix 5B) and $\Pr(X = x|Z = z)^{-1}$ is the treatment weight. An estimator of this weighted risk is $\frac{1}{n}\sum_i \frac{Y_i I(X_i=x)}{\widehat{\Pr}(X_i=x|Z_i)}$ where i indexes the independent and identically distributed n people included in the sample.

$\widehat{\Pr}(X_i = x|Z_i)$ can be estimated nonparametrically or using a parametric model, commonly a logistic regression, called a *propensity score model*.

With missing data

When some data are missing, the identification conditions described above are not sufficient because the weighted risk above is no longer expressed in terms of fully observed data. Let $R = 1$ for complete cases. We can point-identify the risk among the complete cases by incorporating a second weight, formally

$$E\left[\frac{Y\mathrm{I}(X = x)}{\Pr(X = x|Z = Z)}\right] = E\left[\frac{\Pr(R = 1|Z = z, X = x, Y = y)}{\Pr(R = 1|Z = z, X = x, Y = y)}\frac{Y\mathrm{I}(X = x)}{\Pr(X = x|Z = Z)}\right]$$

$$= E\left[\frac{E[I(R = 1)|Z = z, X = x, Y = y]Y\mathrm{I}(X = x)}{\Pr(R = 1|Z = z, X = x, Y = y)\Pr(X = x|Z = z)}\right]$$

$$= E\left[\frac{Y\mathrm{I}(X = x)\mathrm{I}(R = 1)}{\Pr(R = 1|Z = z, X = x, Y = y)\Pr(X = x|Z = z)}\right].$$

The first equality is multiplication by 1, the second is the equivalence between probability and expectation of an indicator function, and the third is the law of total probability. The additional weight, $\Pr(R = 1|Z = z, X = x, Y = y)^{-1}$, is the inverse of the conditional-probability of missingness, hereinafter the *missingness weight*. We require additional conditions to obtain $\Pr(R = 1|Z = z, X = x, Y = y)$ because some data on Z, X, or Y are not observed for people with $R \neq 1$. These conditions are that the data are *missing at random* (MAR) and there is positivity. We reserve full explanation of MAR and estimation of $\Pr(R = 1|Z = z, X = x, Y = y)$ for section 4. In the missing data setting, positivity means that everyone has a non-zero probability of being a complete case.[81,85] Once we obtain the missingness weight, we subsequently obtain the treatment weight among the weighted complete cases (proof in Appendix 5C).[124] An estimator of the weighted risk is

$$\frac{1}{n_{cc}}\sum_i \frac{Y_i I(X_i = x, R_i = 1)}{\widehat{\Pr}(R_i = 1|Z_i = z, X_i = x, Y_i = y)\widehat{\Pr}\,(X_i = x|Z_i = z)},$$

where n_{cc} is the number of complete cases, $n_{cc} = \sum_i I(R_i = 1)$.

Sun and Tchetgen Tchetgen estimators for the missingness weight

To aide illustration, we introduce a simple example with exposure X, outcome Y and confounder Z. There are nonmonotone missing data with four patterns (Table 5.3). Let R denote the pattern to which an individual belongs where $R = 1$ is reserved for complete cases. For the missingness weight, we require the conditional probability of being a complete case, $P(R = 1|Z = z, X = x, Y = y)$, which is the complement of the sum of the probabilities of the other patterns,

$$P(R = 1|Z, X, Y) = 1 - P(R = 2|Z, X, Y) - P(R = 3|Z, X, Y) - P(R = 4|Z, X, Y).$$

The MAR condition is that missingness is independent of the missing data, conditional on the observed data. For example, Y is missing in pattern $R = 2$. Under MAR, $R = 2 \coprod Y|Z, X$, such that $P(R = 2|Z, X, Y) = P(R = 2|Z, X)$. Applying this condition to each pattern, we get

$$\Pr(R = 1|Z, X, Y) = 1 - \Pr(R = 2|Z, X) - \Pr(R = 3|X) - \Pr(R = 4|X, Y).$$

When missingness is not independent from the missing data, neither marginally nor conditional on observed data, then data are *missing not at random* (MNAR). Data are *missing completely at random* (MCAR) when missingness is marginally independent of the observed and missing data, a stronger condition than MAR.

Sun and Tchetgen Tchetgen developed two estimators for the conditional probability of being a complete case under MAR.[81,85] Let π_r be the probability for each pattern. We specify logistic models for each pattern $R > 1$, formally

$$\Pr(R = 2|Z, X) = \pi_2(\gamma_2) = 1/(1 + \exp(-(\gamma_{20} + \gamma_{21} X + \gamma_{22} Z)))$$

$$\Pr(R = 3|X) = \pi_3(\gamma_3) = 1/(1 + \exp(-(\gamma_{30} + \gamma_{31} X)))$$

$$\Pr(R = 4|X, Y) = \pi_4(\gamma_4) = 1/(1 + \exp(-(\gamma_{40} + \gamma_{41} X + \gamma_{42} Y))).$$

The first estimator is the unconstrained maximum likelihood estimator (UMLE). We maximize the joint log-likelihood of the models using the observed data,

$$\ln \mathcal{L}(\gamma) = \sum_{i=1}^{n} \left\{ \left[\sum_{r=2}^{4} I(R_i = r) \ln \pi_r(\gamma_r) \right] + I(R_i = 1) \ln \left[1 - \sum_{r=2}^{4} \pi_r(\gamma_r) \right] \right\}.$$

Each individual contributes a term to the likelihood that corresponds to the pattern to which she belongs. The log-likelihood can be maximized in standard software and we provide code in SAS (using the NLMIXED procedure) and R (using nlm in the Stats package)[125] (Appendix 5D). The UMLE does not naturally impose the constraint that $\pi_1 > 0$, so the log-likelihood may fail to converge if there is a fitted π_1 for a complete case close to zero. Therefore, Sun and Tchetgen Tchetgen also proposed the constrained Bayesian estimator (CBE).

The CBE bounds the probability of being a complete case away from zero by discarding draws from the posterior that do not meet a user-specified constraint. Let c be a user-specified constraint that is a small positive number. CBE produces a posterior distribution that is proportional to the combination of the likelihood, constraint, and prior distributions $f(\gamma)$,

$$f(\gamma|X,Z,Y) \propto \mathcal{L}(\gamma) I(R_i = 1) I(\pi_2(\gamma_2) + \pi_3(\gamma_3) + \pi_4(\gamma_4) < 1 - c) f(\gamma).$$

Sun and Tchetgen Tchetgen used diffuse priors, $\gamma \sim N(0,100)$, $c = 10^{-8}$ and adaptive Gibbs sampling to sample from the posterior.[126] They used the median of the posterior samples to estimate $\hat{\gamma}$. ST provided OpenBugs code for implementation.[127] In Appendix 5D, we provide R code using R2jags package[128] which calls Just Another Gibbs Sampler (JAGS).[129] The OpenBugs and R2jags code are opaque, so we provide code in SAS and R for a more transparent (though less efficient) manually coded Metropolis-Hastings algorithm with rejection sampling that imposes the user-specified constraint by rejecting draws that violate it.[130] Once $\hat{\gamma}$ are estimated by UMLE or CBE, the missingness weight can be estimated for each complete case.

Inference

The naïve standard error estimate from a weighted analysis is not consistent and resulting Wald-type confidence intervals may have poor coverage.[131] There are at least three options for estimating the standard error and obtaining appropriate confidence intervals (CI): 1) "robust" (Huber-White) sandwich estimator, 2) nonparametric bootstrap, or 3) sandwich estimator based on stacked estimating functions. The robust sandwich estimator, which assumes the weights are known (i.e., not estimated), is easy to

implement and expected to produce conservative estimates leading to over-coverage in CIs.[27,132] Nonparametric bootstrapping will produce nominal CIs though it can be computationally intensive.[133,134] Sun and Tchetgen Tchetgen provide stacked estimating functions when the parameter of interest is the conditional odds ratio that will produce nominal CIs, though implementation is less user-friendly compared to the other options. Here we use the robust sandwich estimator.

Simulation

We conducted a limited simulation study to assess finite-sample performance of weighting for nonmonotone missingness. The design was guided by the motivating application. We used both SAS and R.

Data generation

We simulated 5000 studies each with n independent individuals, a binary exposure X, a binary outcome Y, and 3 correlated confounders $\boldsymbol{Z}$ (one continuous and two binary). Under no exposure ($X = 0$), the marginal incidence of the outcome was 10%. See Appendix 5E for data generation details and causal diagram. We varied $n = (1500, 5000)$, marginal prevalence of exposure $p_x = (15\%, 50\%)$, and the true risk difference, in percentage points, $\theta = (0,5)$.

After generating the full data, we induced missingness guided by our motivating example. In the primary scenario, we generated missing data with 6 patterns and 50% complete cases (Table 5.4). For each pattern where $R > 1$, we specified a logistic model

$$\Pr(R = r|Z, X, Y) = 1/\big(1 + \exp\big(-(\gamma_{r0} + \gamma_{r1}X + \gamma_{r2}Y + \gamma_{r3}Z_1 + \gamma_{r4}Z_2 + \gamma_{r5}Z_3)\big)\big)$$

to obtain individual probabilities of being in that pattern. The probability of being a complete case was the complement of the sum of the other probabilities. The observed missing data pattern, R, was generated from a multinomial distribution and then missing data were imposed according to the observed pattern (i.e., value of any variable missing under that pattern was set to missing). We varied the γ coefficients to produce missing data that were MAR (coefficients for variables missing in that pattern were zero), MCAR (all coefficients were zero), and MNAR (some coefficients for missing variables were non-zero).

The γ intercepts were set by numerical approximation to achieve the desired prevalence of each pattern.[135] We also explored how the percent of complete cases and number of patterns affected results when data are MAR (secondary scenarios, Table 5.4).

Analysis

In each simulated cohort, we implemented three approaches to address missing data. Regardless of missing data approach used, we used inverse probability of treatment weighting to address confounding and fit a weighted linear-binomial outcome model conditional on exposure ($E(Y) = \lambda + \theta X$) using generalized estimating equations with an independence covariance structure to estimate the risk difference, θ, and robust standard error for Wald-type 95% CIs.

First, we conducted an analysis restricted to complete cases (i.e., conditional on $R = 1$). We fit the treatment propensity score model among the complete cases and estimated the treatment weight as $\widehat{\Pr}(X = x|Z = z, R = 1)^{-1}$.

Second, we implemented weighting for missingness. We implemented UMLE and, if UMLE failed to converge, we estimated $\hat{\gamma}$ as the posterior median obtained by CBE implemented by adaptive Gibbs sampling with a single chain of 10,000 iterations with 5,000 burn-in samples discarded. We used diffuse priors $\gamma \sim N(0,100)$ and set $c = 10^{-8}$. For each missingness model, all variables observed in that pattern were included. Models were correctly specified except in MNAR scenarios. Using $\hat{\gamma}$, we estimated the missingness weight, $\widehat{\Pr}(R = 1|Z = z, X = x, Y = y)^{-1}$. We subsequently fit the missingness-weighted treatment propensity score model among the complete cases to obtain the treatment weight, $\widehat{\Pr}(X = x|Z = z)^{-1}$. The final weight was the product of the missingness weight and treatment weight.

Third, we implemented MI by chained equations.[136,137] Using logistic regression for binary variables and linear regression for continuous variables, we imputed missing data 20 times (mice package in R[138] and MI procedure in SAS). All variables were included in each imputation model which were correctly specified except in MNAR scenarios. In each imputed dataset, we fit the treatment propensity

score model and estimated the treatment weight, $\widehat{\Pr}(X = x|Z = z)^{-1}$. From the treatment-weighted outcome models, the 20 estimates were combined by Rubin's rule,[139] ($\bar{\theta} = \frac{1}{20}\sum_{k=1}^{20}\hat{\theta}_k$ where $\hat{\theta}_k$ is the estimated risk difference from imputation k; $V(\bar{\theta}) = \frac{1}{20}\sum_{k=1}^{20}\hat{V}(\hat{\theta}_k) + \left(1 + \frac{1}{20}\right)\left(\frac{1}{20-1}\right)\sum_{k=1}^{20}(\hat{\theta}_k - \bar{\theta})^2$ where $\hat{V}(\hat{\theta}_k)$ is the estimated robust variance for $\hat{\theta}_k$ in imputation k).

Finally, we analyzed the full data without inducing missingness. While not available in practice, the full data provide a reference against which we compare the three approaches. We fit the treatment propensity score model in the full data and estimated the treatment weight as $\widehat{\Pr}(X = x|Z = z)^{-1}$.

To compare estimator performance, we estimated 1) number of times the estimator failed to produce results, 2) bias (mean of the risk difference estimates minus the true risk difference), 3) empirical standard error (standard deviation of the estimates), 4) root mean squared error (square root of the average of the squared differences between each risk difference estimate and the true risk difference), 5) average model standard error (square root of the average of the variance estimates), and 6) CI coverage (proportion of estimated 95% CIs that included the true risk difference).[117] Failures did not contribute to the estimation of the other performance measures. We also captured the run time as the average time in seconds for a single cohort (each scenario was run separately on a single 2.5GHz processor with up to 15GB of memory allocated; versions R 4.1.0 and SAS 9.4).

Results

Failure to produce results

For all approaches, failures were rare and occurred with similar frequency across the missing data approaches (Table 5.5). In the primary scenario, failures only occurred when exposure prevalence was 15%, sample size was 1500, and data were missing MAR (≤0.5% failures) or MNAR (≤4.0%). In the primary scenario, there were no failures at the larger sample size, 50% exposure prevalence, or when data were MCAR. In general, failures for weighting and complete analysis occurred when the weighted outcome model restricted to the complete cases did not converge due to too few exposed individuals or too few outcomes. Weighting can also fail if UMLE fails to converge and thus missingness weights are

not estimate, however this UMLE non-convergence never occurred in SAS and occurred just twice in R (across 5000 cohorts in 44 scenarios). In those 2 instances, CBE successfully estimated the weights. For the remainder of the simulation results, only UMLE is presented for weighting.

Statistical performance

Results from R and SAS were approximately equivalent; we present R only. For the primary scenario (6 patterns, 50% complete cases), results are presented in Table 5.6 and Figure 5.1 (MAR), Appendix Table 5F.1 and Appendix Figure 5F.1 (MCAR), and Appendix Table 5F.2 and Appendix Figure 5F.2 (MNAR). When data were MAR, complete case analysis was notably biased and thus had poor coverage. At 50% exposure prevalence, MI and weighting performed similarly with negligible bias and nearly the same RMSE and coverage. At 15% exposure prevalence with n=1500, weighting performance declined with a small increase in bias and reduced precision and coverage. Comparably, MI had greater precision and thus lower RMSE. When data were MCAR, bias was negligible for all approaches. Weighting had the same precision as complete case analysis. Again, improved precision of MI over weighting was apparent only at 15% exposure prevalence with n=1500. When data were MNAR, all missing data approaches were biased. At 15% exposure prevalence, weighting had more bias than MI though at 50% exposure prevalence bias was similar.

Results from scenarios varying percent of complete cases and number of missing data patterns are presented in Appendix Tables 5F.3 and 5F.4. Data were MAR with 50% exposure prevalence. Bias was negligible for MI and weighting so plots of results (Figure 5.2 and Appendix Figure 5F.3) focus on RMSE and coverage. At n=1500, MI had a lower RMSE than weighting when there were 35% complete cases. A small difference persisted at 50% complete cases only when there were 8 patterns and the difference nearly disappeared at 65% complete cases. A similar pattern was present when n=5000 though the differences between MI and weighting were smaller. For both sample sizes, there was decreasing coverage for MI (dropping <95%) and increasing coverage for weighting (>95%) as the percent of complete cases declines. There were little differences in results when there were 6 or 8 patterns.

Computational efficiency

Figure 5.3 plots the time in seconds for MI and weighting for all 22 scenarios by sample size (Appendix Table 5F.5 for numeric results). On average, MI took 2.7 times as long as weighting in R (13.0 vs. 4.9 seconds) and 18.3 times as long in SAS (19.2 vs. 1.0 seconds). Weighting was faster in SAS than in R (average 1.0 vs. 4.9); MI was slower in SAS than in R (average 19.2 vs.13.0).

Application

Methods

As outlined above, we aimed to estimate the effect of maternal anemia in pregnancy on the risk of spontaneous preterm birth. We implemented three approaches to address missing data that mirror the simulation study. In all, we used treatments weights to address confounding by gestational age at enrollment, maternal age, maternal HIV serostatus, and birth history including parity, prior preterm birth, and prior stillbirth. We fit a weighted linear-binomial model to estimate the risk difference and obtained Wald-type 95% CIs using the robust standard error. Gestational age and maternal age were modeled with restricted quadratic splines with 4 knots at the 5th, 35th, 65th, and 95th percentiles.[118]

First, we conducted a complete case analysis restricted to the 781 people with complete data. Second, we implemented weighting for missingness. There were a number of missing data patterns with few people so we combined rare patterns as suggested by Sun and Tchetgen Tchetgen.[85] We combined patterns in two ways: 1) patterns 8 through 16 (each <1%, total 2.4%) resulting in 8 total patterns and 2) patterns 6 through 16 (each <2%, total 5.7%) resulting in 6 total patterns. All observed variables were included in the missingness models. For the combined pattern, the model included variables observed across all patterns in the set: gestational age, nulliparity, and prior preterm birth. We implemented UMLE and CBE (with diffuse priors $\gamma \sim N(0,100)$ and $c = 10^{-8}$). We used Gelman-Rubin statistic (Rhat) to assess convergence.[140] We estimated $\hat{\gamma}$ by the median across 3 chains of 120,000 iterations with the first half of samples discarded. Using $\hat{\gamma}$, we estimated the missingness weights for the complete cases and then fit the missingness-weighted treatment propensity score model to estimate the treatment weights. The final weight was the product of the two weights.

Third, we implemented MI by chained equations. We implemented MI four ways varying 1) whether spline terms were created before imputation (active imputation) or after imputation (passive imputation) and 2) imputation model flexibility. The imputation models included all variables from the analysis and the exposure-outcome interaction. To increase flexibility, we included interactions between the outcome and strong outcome predictors: prior preterm birth, maternal age, and HIV serostatus. We imputed missing data 20 times. In each imputed dataset, we fit the treatment propensity score model, estimated the treatment weights and then estimated the risk difference and robust variance from the treatment-weighted linear-binomial model. The 20 estimates were combined by Rubin's rule (as detailed above).

Results

In the complete case analysis, the estimate of the effect of anemia on spontaneous preterm birth was 2.7 percentage points (95% CI -5.7, 11.0) (Table 5.7). Although anemia appeared to elevate the risk of preterm birth, the confidence interval was wide and encompassed effects ranging from strongly protective to strongly causative. While addressing missingness resulted in small changes in the point estimate, the confidence intervals remained wide so that these changes were not meaningful. Across the weighting approaches, the point estimates varied from 2.1 to 2.3. There was modest improvement in statistical efficiency with weighting compared to complete case analysis (standard error: complete case 4.3, UMLE 4.0, CBE 4.1). Across the MI specifications, the point estimates varied more widely, from 1.9 to 3.0. The standard errors ranged from 3.8 to 4.1.

Discussion

Our work supports the finding of Sun and Tchetgen Tchetgen that weighting is an alternative to MI to account for nonmonotone missingness. The UMLE estimator for missingness weights is easy to implement in commonly used software and intuitive, in contrast to prior approaches for weighting for nonmonotone missingness.[81,83–85] We provide code in R and SAS to support uptake of this approach.

Although theoretically UMLE may fail to converge if there are complete cases with a small probability of being a complete case, this was extremely rare in our simulation study (twice in 220,000

simulated cohorts). CBE is an alternative that can be used in these settings, though implementation is more challenging. Future work could examine characteristics of the datasets in which UMLE failed in an effort to anticipate failures. Even when UMLE converges, the weighted outcome model may fail, though, this was also rare in our simulation study (≤0.5% under MAR). Weighting failed when the complete case analysis also failed as both approaches fit the outcome model restricted to the complete cases. The complete case analysis could be used to assess whether weighting is likely to succeed. UMLE non-convergence was previously examined in a simple simulation by Sun and Tchetgen Tchetgen.[85] They reported that UMLE failed in approximately 1% of simulation replicates. However, the authors did not say specifically whether these failures were due to UMLE non-convergence or non-convergence of the weighted outcome model. They also do not report failures of the other approaches.

Generally weighting is a semi-parametric approach; in contrast, the more widely-used MI is a fully parametric approach. The stronger parametric assumption improves precision of MI,[79] however, in our simulation, this improved efficiency was only notable at the smaller sample size with 15% exposure prevalence. This observed relationship with exposure prevalence may be because there was a lot of missingness in the exposure in our simulation (30% in primary scenario marginally). In the other scenarios, there was little practical difference in efficiency between weighting and MI. To improve efficiency of weighting, augment weighted estimators (also derived by Sun and Tchetgen Tchetgen) could to be used.[85] We chose to use the non-augmented estimators in this work because they are easier to implement and thus more likely to be used in practice.

There were two instances where performance was notably different between weighting and MI. First, when data were MNAR, both MI and weighting were biased, however MI was less biased than weighting. It is not clear whether this is a characteristic of the estimators themselves or a produce of our data generating mechanism. In Sun and Tchetgen Tchetgen's simulation, there was not a consistent pattern of bias in the estimators across the parameters of interest under MNAR.[85] Second, weighting was computationally faster than MI. At the small sample sizes examined here, the difference was trivial, however with large sample sizes or when estimating the variance using resampling algorithms with a

moderate sample size, the difference could be meaningful. Under linear extrapolation of our results, 1000 bootstrap resamples for n=5000 is expected to take 7.8 hours using MI and 0.4 hours using weighting in SAS. In R, MI is expected to take 5.1 hours and weighting 1.9 hours.

In our empirical example, there was little change in the estimates across the approaches and results were imprecise. The application highlights that implementing weighting and MI each require a number of analytic decisions. For weighting, some missing data patterns were rare, so we collapsed rare patterns. The model for this combined set could only include variables that were observed across all patterns in the set and therefore induced an assumption stronger than MAR. Combining patterns may induce bias, but such bias is likely small given that the combined patterns include few observations.[85] After estimation, it is good practice to examine the distribution of the weights as extreme weights can inflate the variance. Truncation of extreme weights can improve precision though potentially at the cost of bias.[141] Arguably, MI demands a greater number of analytic decisions.[79,142] Some of these decisions include: the iterative procedure (Markov chain Monte Carlo vs. chained equations); transforming skewed variables; passive vs. active imputation; number of iterations and imputed datasets; and specifying prior distributions.[78,80,142–151]

Both weighting and MI require correct model specification. In MI, we specify a model for each variable with missingness. In weighting, we specify a model for each pattern. The number of potential patterns grows exponentially with the number of variables with missingness meaning weighting will often require specification of more models than MI. The missingness models in weighting all have binomial dependent variables whereas the distribution of the dependent variables in the imputation models may take different distributions. Some imputation approaches assume a normal distribution and if this assumption is grossly violated, performance can be poor.[149] It may be necessary to transform variables before imputation. Although it has been argued that it is easier to correctly specify the missingness models than the imputation models,[79] accurate specification of models depends on context-specific knowledge of what variables cause missingness and the functional form of those variables with missingness itself (for missingness models) or with the unobserved data (for imputation models). Finally,

an important limitation of MI is that specification of the imputation models can impose constraints on the model of interest (i.e., the analyst must ensure the imputation models are *congenial* with the analysis model).[147] Issues of congeniality can arise by omitting the outcome or important interactions from imputation models, or using different functional forms of variables in imputation models and analysis models. This is not the case with weighting where specification of the missingness models is independent of the full data and therefore does not impose restrictions on the analysis model.

Our work has limitations. Our simulation design was relatively simple compared to most data settings and we did not vary all elements of the data generating mechanism. We mimicked the patterns observed in the application where missing exposure and outcome accounted for most of the missingness and results may differ as these patterns change. Finally, other than the MCAR and MNAR scenarios, we did not vary the strength of relationships between observed data and missingness. A strength of our work is that our simulation employed increased complexity over the only other simulation examining these estimators.[85] Additionally, in contrast to that previous simulation, we focused on estimating marginal effects and combined missing data approaches with inverse probability of treatment weighting to address confounding. We also provide code for the estimators in two commonly used software programs.

Conclusion

Weighting can be used to address confounding and missing data simultaneously.[124] When missingness is nonmonotone, Sun and Tchetgen Tchetgen's UMLE is an easy to implement estimator for the missingness weights. Though less statistically efficient in some settings, weighting is a viable alternative to MI. Weighting for missingness may be more intuitive for researchers already familiar with using weighting to address other biases and computational efficiency is attractive in large datasets or when estimating the variance using resampling algorithms such as nonparametric bootstrap.

Tables

Table 5.1. Cohort characteristics overall and stratified by anemia status at enrollment.

	N (%)			
		Anemia status		
Characteristic	Overall (N=1447)	Positive (N=138)	Negative (N=884)	Unknown (N=425)
Spontaneous preterm birth	120 (9.9)	15 (12.4)	70 (9.6)	35 (9.7)
Missing	239	17	157	65
At enrollment				
Gestational age (weeks)				
Median (IQR)	16.1 (13.3, 18.3)	16.9 (14.7, 18.9)	15.9 (12.9, 18.1)	16.6 (13.6, 18.6)
First trimester	432 (29.9)	29 (21)	286 (32.4)	117 (27.5)
Missing	0	0	0	0
Maternal age				
Median (IQR)	27 (23, 32)	27 (22, 33)	27 (23, 32)	27 (22, 31)
<20 years	111 (7.9)	11 (8.3)	59 (6.9)	41 (9.8)
20-34 years	1113 (79.2)	102 (76.7)	677 (79.2)	334 (79.9)
≥35 years	182 (12.9)	20 (15)	119 (13.9)	43 (10.3)
Missing	41	5	29	7
HIV+	349 (24.2)	58 (42)	195 (22.1)	96 (22.7)
Missing	3	0	0	3
Nulliparous	457 (31.6)	40 (29)	286 (32.4)	131 (30.8)
Missing	0	0	0	0
Prior preterm birth	410 (28.3)	40 (29)	258 (29.2)	112 (26.4)
Missing	0	0	0	0
Prior stillbirth	126 (9.3)	15 (11.6)	80 (9.6)	31 (7.8)
Missing	85	9	48	28

Abbreviations: IQR, interquartile range

Table 5.2. Missing data patterns in the motivating example data from the ZAPPS cohort, n=1447.

Pattern	Anemia	Spont. PTB	Maternal Age	HIV serostatus	Prior stillbirth	N	%
1	O	O	O	O	O	781	54.0
2	M	O	O	O	O	330	22.8
3	O	M	O	O	O	151	10.4
4	M	M	O	O	O	58	4.0
5	O	O	O	O	M	45	3.1
6	M	O	O	O	M	26	1.8
7	O	O	M	O	O	21	1.5
8	O	M	M	O	O	12	0.8
9	O	M	O	O	M	11	0.8
10	M	M	M	O	O	5	0.4
11	M	O	O	M	O	2	0.1
12	O	O	M	O	M	1	0.1
13	M	O	M	O	O	1	0.1
14	M	O	M	O	M	1	0.1
15	M	M	O	O	M	1	0.1
16	M	M	O	M	O	1	0.1

"O" indicates variable is observed and "M" indicates variable is missing
Variables not included here (gestational age at enrollment, maternal HIV serostatus, nulliparity, and prior preterm birth) did not have missingness
Abbreviations: Spont. PTB, spontaneous preterm birth

Table 5.3. Missing data patterns in the simple illustrative example.

Pattern (R)	Z	X	Y
1	O	O	O
2	O	O	M
3	M	O	M
4	M	O	O

"O" indicates variable is observed and "M" indicates variable is missing

Table 5.4. Missing data patterns in the simulation study.

						% in each pattern					
Pattern (R)	X	Y	Z_1	Z_2	Z_3	Primary	Secondary scenarios				
1	O	O	O	O	O	50	65	35	50	65	35
2	M	O	O	O	O	15	10	15	10	5	15
3	O	M	O	O	O	15	10	15	10	5	10
4	M	M	O	O	O	10	5	15	10	5	10
5	O	O	O	O	M	5	5	10	5	5	10
6	M	O	O	O	M	5	5	10	5	5	10
7	O	O	M	O	O				5	5	5
8	O	M	M	O	O				5	5	5

"O" indicates variable is observed and "M" indicates variable is missing

Table 5.5. Failures in 5000 simulated datasets when exposure prevalence was 15% and sample size was 1500 for primary missing data scenario (6 patterns, 50% complete cases.

	Risk difference 0%		Risk difference 5%	
Missing data approach[a]	R	SAS	R	SAS
MAR				
CC	12 (0.2%)	25 (0.5%)	0 (0.0%)	1 (<0.1%)
MI	18 (0.4%)	25 (0.5%)	0 (0.0%)	1 (<0.1%)
Weighting[b]	12 (0.2%)	25 (0.5%)	0 (0.0%)	1 (<0.1%)
MNAR				
CC	165 (3.3%)	202 (4.0%)	32 (0.6%)	52 (1.0%)
MI	163 (3.3%)	168 (3.4%)	22 (0.4%)	38 (0.8%)
Weighting[b]	162 (3.2%)	202 (4.0%)	34 (0.7%)	52 (1.0%)

Abbreviations: MAR, missing at random; CC, complete case analysis; MI, multiple imputation; MNAR, missing not at random

[a]All approaches addressed confounding using inverse probability of treatment weights

[b]UMLE used to estimate the missingness weights

Table 5.6. Bias, empirical standard error, root mean squared error, average model standard error, and 95% confidence interval coverage of the risk difference (in percentage points) for primary missing data scenario (6 patterns, 50% complete cases) when data are missing at random (MAR).

	Exposure prevalence 15%					Exposure prevalence 50%				
Missing data approach[a]	Bias	ESE	RMSE	avg. ModSE	95% CI Coverage	Bias	ESE	RMSE	avg. ModSE	95% CI Coverage
Risk difference 0										
n=1500										
Full	-0.1	2.2	2.2	2.2	94%	0.0	1.6	1.6	1.6	96%
CC[b]	-1.6	2.8	3.2	2.8	81%	-1.1	1.9	2.2	1.9	91%
MI[b]	0.4	4.0	4.1	4.3	94%	0.0	2.6	2.6	2.6	94%
Weighting[b,c]	-0.5	4.4	4.5	4.5	90%	0.0	2.7	2.7	2.7	95%
n=5000										
Full	0.0	1.2	1.2	1.2	95%	0.0	0.9	0.9	0.9	95%
CC	-1.6	1.5	2.2	1.5	77%	-1.0	1.0	1.4	1.0	84%
MI	0.2	2.3	2.3	2.4	94%	0.1	1.4	1.4	1.4	95%
Weighting[c]	-0.2	2.4	2.4	2.5	94%	0.1	1.4	1.4	1.5	96%
Risk difference 0.05										
n=1500										
Full	-0.1	2.6	2.6	2.6	95%	0.0	1.7	1.7	1.8	95%
CC	-3.4	3.5	4.9	3.5	74%	-2.5	2.1	3.3	2.2	78%
MI	0.0	4.7	4.7	4.8	94%	0.0	2.8	2.8	2.8	94%
Weighting[c]	-0.6	5.2	5.2	5.3	91%	0.0	2.8	2.8	3.0	96%
n=5000										
Full	0.0	1.4	1.4	1.4	95%	0.0	1.0	1.0	1.0	95%
CC	-3.4	1.9	3.8	1.9	55%	-2.5	1.2	2.7	1.2	45%
MI	-0.1	2.6	2.6	2.6	94%	0.1	1.5	1.5	1.5	95%
Weighting[c]	-0.1	2.8	2.8	2.9	95%	0.1	1.5	1.5	1.6	96%

Abbreviations: CC, complete case analysis; MI, multiple imputation; ESE, empirical standard error; RMSE, root mean squared error; avg. ModSE, average model standard error; CI, confidence interval

[a]All approaches addressed confounding using inverse probability of treatment weights

Results from 5000 simulated cohorts except approaches marked with [b] at 15% prevalence (see Table 5 for number of failures)

[c]UMLE used to estimate the missingness weights

Table 5.7. Risk difference estimates (in percentage points) and uncertainty from application examining effect of anemia on spontaneous preterm birth.

Missing data approach[a]	RD	95% CI[b]	SE[b]
Complete case analysis	2.7	-5.7, 11.0	4.3
Weighting, UMLE, 6 patterns[c]	2.1	-5.8, 10.0	4.0
Weighting, UMLE, 8 patterns[d]	2.0	-5.9, 9.9	4.0
Weighting, CBE, 6 patterns[c]	2.3	-5.8, 10.3	4.1
Weighting, CBE, 8 patterns[d]	2.2	-5.8, 10.2	4.1
MI, transform/impute[e]	1.9	-5.7, 9.6	3.8
MI, transform/impute, more flexible[e,f]	3.0	-5.3, 11.2	4.1
MI, impute/transform[g]	2.4	-5.6, 10.5	4.0
MI, impute/transform, more flexible[g,f]	3.0	-5.0, 11.0	4.0

Abbreviations: RD, risk difference in percentage points; CI, confidence interval; SE, standard error; UMLE, unconstrainted maximum likelihood estimator; CBE, constrained Bayesian estimator; MI, multiple imputation

[a]All approaches addressed confounding using inverse probability of treatment weights

[b]Robust standard error estimated by "robust" (Huber-White) sandwich estimator

[c]Combined patterns 6 through 16 (each <2%) resulting in 6 total patterns

[d]Combined patterns 8 through 16 (each <1%) resulting in 8 total patterns

[e]Spline terms were created first and then missing data were imputed

[f]More flexible imputation models included interactions between outcome and strong predictors of the outcome: prior preterm birth, maternal age, and HIV serostatus

[g]Missing data were imputed first and then spline terms were created

Figures

Figure 5.1. Boxplots of risk difference estimates for primary missing data scenario (6 patterns, 50% complete cases) when data are missing at random (MAR). Panel A) true risk difference is 0%; panel B) true risk difference is 5%. Horizontal black line marks true risk difference; black dot marks mean; small gray dots are a 10% random sample of estimates. Abbreviations: CC, complete case analysis; MI, multiple imputation

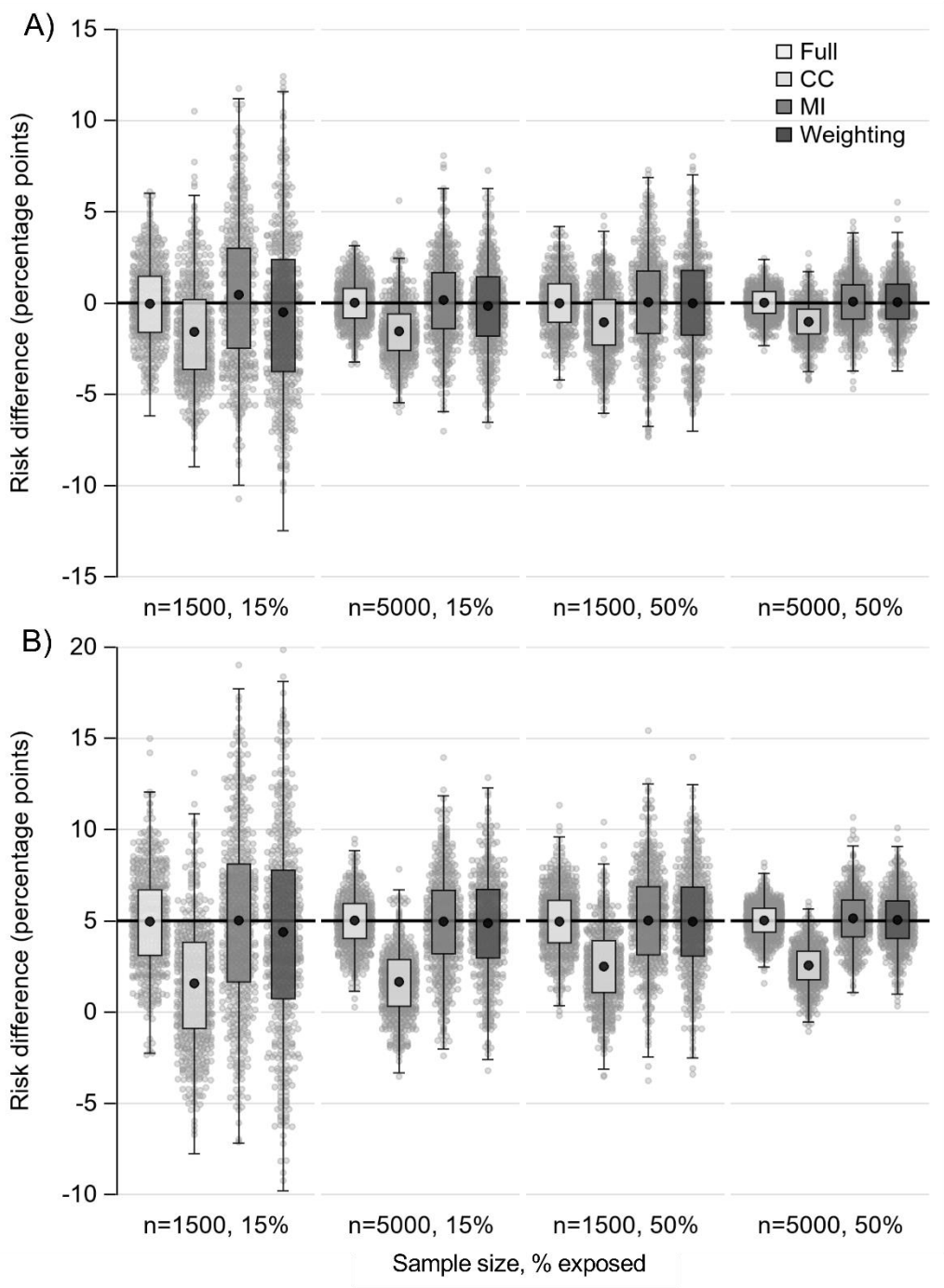


Figure 5.2. Root mean squared error and confidence interval coverage as percent complete cases and number of patterns varies when the true risk difference is 5% and data are missing at random (MAR). Solid line indicates 6 patterns and dashed line indicates 8 patterns (for Full, the two lines are exactly overlaid). Abbreviations: MI, multiple imputation; CI, confidence interval

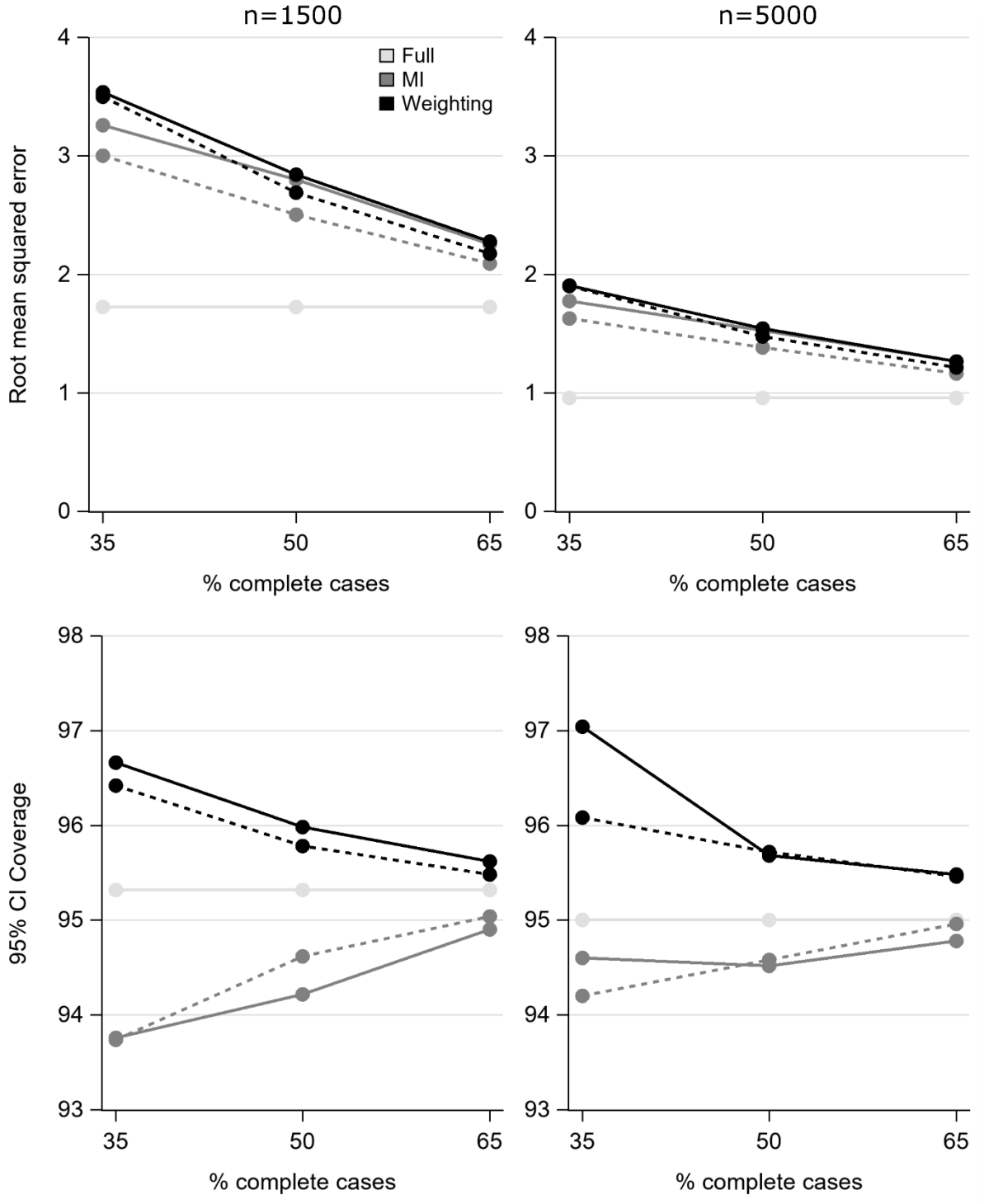
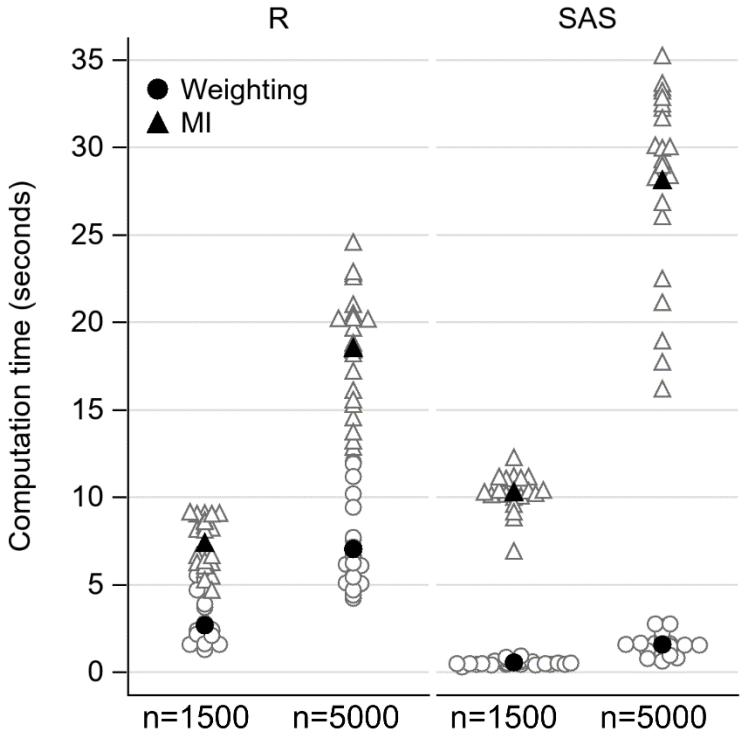


Figure 5.3. Computational time in seconds to implement MI (triangle) and weighting using UMLE (circle) from 22 scenarios by sample size and computer program. Filled in black symbol marks the mean. Abbreviations: MI, multiple imputation; UMLE, unconstrained maximum likelihood estimator.



CHAPTER 6: DISCUSSION

Overview of key findings

In this work we examined approaches to address missing data and measurement error in observational epidemiologic studies. This methodological work was motivated by research aimed at understanding the underlying causes of preterm birth and identifying points of potential intervention to prevent preterm birth.

Aim 1

Outcome misclassification can produce meaningful bias in estimates of risk and causal effects. We introduced standardization estimators that leverage external validation data to account for outcome misclassification. Our simulation showed that these estimators were unbiased for the marginal risk under the natural course and causal effects when assumptions held.

To accommodate continuous variables or high dimensional data, we relied on parametric modeling. To estimate the conditional outcome risk, we used a modified likelihood previously described by Carroll et al. and Lyles et al.[61,101] Lyles et al. also described using logistic regression to estimate flexible misclassification models in validation data. We have incorporated this flexibility into our estimators as well. In Lyles et al., the parameters of interest were the coefficients a logistic regression for the outcome. In our work, the parameters of the outcome logistic model were nuisance parameters; rather we were interested in estimating marginal effects. Thus, we added a standardization step to the estimators described in Lyles et al. With this standardization, the variance estimation proposed in Lyles et al. is no longer valid. Therefore, we used M-estimation and the empirical sandwich variance estimator.[115,116] Our simulation showed that the empirical sandwich variance estimator appropriately estimated the empirical standard error.

Because we are leveraging external validation data, we need to transport the misclassification model from the validation data to the main study.[61(sec2.2.4)] We rely on an ignorability assumption for transport. For this assumption, we condition on variables related to misclassification and whose distribution varies between the validation data and the main study such that, conditional on these variables, inclusion in each population is independent of the misclassification model (mismeasured outcome conditional on true outcome). We introduce two estimators that account for these variables. Versions of these estimators (in the simpler setting without exposure or confounders) were previously introduced in Edwards et al. and Ackerman et al.[66,67] Although we focused on external validation data, our conclusions also apply to internal validation data that is not a conditional random sample from the main study, conditional on exposure and confounders.

Our work highlighted that we generally need rich validation data to account for measurement error unless we make strong assumptions such as the error is nondifferential or that the validation data are a random sample of the main study conditional on exposure and confounders.

Aim 2

Weighting is an alternative to MI to account for missing data, even when missingness is nonmonotone. Previous weighting approaches for nonmonotone missingness were cumbersome, but Sun and Tchetgen Tchetgen's novel UMLE approach is straightforward to implement. Sun and Tchetgen Tchetgen's CBE approach is less straightforward, however, in our simulation UMLE rarely failed, making implementation of CBE unnecessary.

In our simulation, performance of weighting and MI were mostly similar, however there were three differences to highlight. As expected, MI was more precise than weighting; though this improvement was notable only at the smaller sample with 15% exposure prevalence. This gain in precision comes from stronger parametric assumptions in MI than in weighting.[79] Also as expected, we observed improved computational efficiency of weighting over MI. Finally, we observed less bias in MI than in weighting when the MAR assumption was violated (i.e., data were MNAR). We do not believe

this is an inherent characteristic of these estimators, however, performance when assumptions are violated is an area for future research.

Ultimately, a number of factors may be considered when choosing between weighting and MI to account for nonmonotone missingness.

1) Model specification: For weighting, we need to correctly specify a model for each missingness pattern. Each model has a binary outcome so the choice of parametric family, binomial, is correct. However, we need to choose the dependent variables with the correct functional form to satisfy our MAR assumption. Specification of these missingness models is independent of the full data and therefore does not impose restrictions on the analysis model. For MI, we need to correctly specify a model for each variable with missingness. The distribution of these variables may vary and a parametric family is chosen for each; there is the possibility that the parametric family may be misspecified. It may be necessary to transform variables if the distribution, e.g., normal, is grossly violated. Additionally, we need to choose the dependent variables with the correct functional form to satisfy our MAR assumption. Importantly, unlike with weighting, specification of the imputation models in MI is not independent of the full data and these imputation models can imposed restrictions on the analysis model. Therefore, imputation models and analysis model must be congenial, meaning both models are compatible with a shared larger model because dependencies excluded from the imputation model restrict the analysis model.[147] This means that in practice the imputation model should be more flexible than the analysis model.

   Although it has been argued that it is easier to correctly specify the missingness models than the imputation models,[79] accurate specification of models depends on context-specific knowledge of what variables cause missingness and the functional form of those variables with missingness itself (for missingness models) or with the unobserved data (for imputation models).

2) Precision: MI is expected to be more precise than weighting. Our simulation illustrated that this improved precision may not always be meaningful. Precision is likely affected by sample size,

exposure prevalence, outcome incidence, the number of missingness patterns, and the number of complete cases.

3) Computational time: Given the relative difference in computation time observed in our simulation, weighting may be preferred some settings where computational efficiency matters, such as when there is a large sample size or when resampling approaches, such as nonparametric bootstrap, are needed to estimate confidence intervals.
4) Comfort: A researcher's comfort level with each approach (and the various decisions required for implementation) is important for correct implementation, interpretation, and communication of results.

Finally, a researcher can implement both approaches. As each approach relies on correct specification of different models, agreement between the approaches may indicate minimal bias.

Limitations

Limitations specific to each aim are presented in the Discussion sections of Chapters 4 and 5. In this section we highlight limitations shared by the two aims.

We performed simulation studies in each aim. In Aim 1, the objective was to illustrative implementation of the proposed estimators, that validity aligned with expectations given the data generation and assumptions of each estimator, and that the empirical sandwich variance estimator was valid. In Aim 2, the objective was to compare performance of weighting to MI to address nonmonotone missing data. We specifically examined the setting of a time-fixed exposure in which weighting was used to account for confounding. A limitation of our work is that these simulation studies were restricted to a small number of settings. As is generally true of simulation studies, we could not and did not aim to examine performance of these estimators across many realistic settings. For example, we did not examine performance when parametric models were misspecified. Additionally, data generation closely mimicked the applied examples and did not dramatically alter data generation in different scenarios, beyond a limited set of factors in Aim 2. The nature of these simulation studies limits the generalizability of our

results. For example, we cannot say whether our findings of the comparative performance of weighting and MI in our second Aim 2 extend to settings with a time-to-event outcome.

In both aims, our parameters of interest included causal effects. To connect the observed outcomes to potential outcomes, we invoked the causal consistency assumption. This assumption is that the observed outcome is the potential outcome if the exposure had been set to the observed exposure value. Formally, for a binary exposure, this can be written as $Y = AY^1 + (1 - A)Y^0$ where Y^a is the potential outcome when A is set to a.[152] This assumption is sometimes combined with an assumption of no interference and collectively called the stable unit-treatment value assumption, referred to as SUTVA.[152] Causal consistency is met by design and is intuitive in a trial where a specific intervention is randomized. This is not the case however in observational studies, particularly those of biologic features.[108] In our applied examples, we examined the exposures of maternal HIV infection and maternal anemia. It is challenging to conceive of interventions that eliminate anemia or HIV infection. Even if we can conceive of these interventions, it is likely such interventions would affect pregnancy outcomes in ways unrelated to anemia or HIV infection. Therefore, the causal consistency assumption is unlikely to hold in our applied examples. If this assumption is violated, then we cannot interpret our estimates as causal effects.

Implications

Epidemiology training and research largely focus on methods for mitigating confounding bias under the strong assumptions of no missing data and no measurement error. Yet, missing data and measurement error are ubiquitous, particularly in observational data sources, and they hamper our ability to produce accurate results. Unfortunately, there are barriers to widespread adoption of principled approaches for missing data and measurement error in applied research, including poor dissemination to broad non-statistical audiences. The research in this dissertation addresses such methodological challenges by developing and applying modern epidemiologic tools with the goal of making them more accessible to researchers. This research supports broader use of these tools to increase the quality of

research and produce more actionable evidence to improve pregnancy outcomes and public health more broadly.

There are four important strengths of this research. First, we focused on epidemiologic tools that address multiple biases. In Aim 1, the estimators address confounding and outcome misclassification. In Aim 2, the estimators address confounding and missing data. Methods research often examines performance of analytic approaches that address one bias in isolation, however this does not reflect realistic data analysis of observational data. Second, we focused on estimators of marginal standardized measures in contrast to previous work. For both aims, the estimators built on prior work that focused on estimation of conditional log odds ratios, i.e., the coefficients in a logistic regression.[85,101] We incorporated modern standardization approaches to estimate marginal measures, g-computation in Aim 1 and weighting in Aim 2. Third, we considered estimators of causal effects. We incorporated potential outcomes notation and relied on a set of causal identification conditions that included conditional exchangeability with positivity and causal consistency. Finally, we provide code to support uptake of these estimators.

To conclude, measurement error and missing data pose serious threats to valid estimation of descriptive and causal parameters. Qualitative conjectures about the direction of potential bias do not produce more accurate results or inform inference. In Aim 1, to address outcome misclassification and confounding, we proposed estimators that leverage external validation data. To appropriately address misclassification and satisfy identification assumptions, we generally need rich and large validation data. The estimators introduced here are able to leverage external already existing data for validation. In Aim 2, we showed that weighting can be used to address confounding and missing data simultaneously.[124] When missingness is nonmonotone, Sun and Tchetgen Tchetgen's UMLE is an easy to implement estimator for the missingness weights. Though less statistically efficient in some settings, weighting is a viable alternative to MI. Weighting for missingness may be more intuitive for researchers already familiar with using weighting to address other biases and computational efficiency is attractive in large datasets or when estimating the variance using resampling algorithms such as nonparametric bootstrap.

Future work will build on this research in several ways. A strength of the current work is that we focused on methodological tools that address multiple biases, measurement error and confounding in Aim 1 and missing data and confounding in Aim 2. However, the analysis of real data often must address all three – missing data, measurement error, and confounding – and potentially more. Future work will examine approaches to address all three biases. One area of research will be to examine the impacts of missing data in our measurement error approaches and how missing data can be addressed. It will be important to distinguish missing data in the study sample and in the external validation sample because the latter is not typically part of the target population. Another area of research will consider how to improve efficiency of approaches that address multiple biases, particularly measurement error. We observed a notable loss of precision when we relaxed our assumption of nondifferential error, particularly for the causal effect estimate. It is important that we allow for differential error as we saw that erroneously assuming nondifferential error can be more biased than the naïve analysis, however relaxing this assumption may produce results that are so imprecise as to be uninformative. Finally, future work will focus on the development and communication of generalizability and transportability methods. Data fusion, the combining of multiple data sources, continues to increase in our field and data fusion necessarily relies on transporting parameters between samples. Further, most modern epidemiology approaches, in general, generalize or transport estimates whether they address confounding, selection bias, or other biases. For example, the average treatment effect involves generalizing risks among the treated and untreated to the total population. Analogously, the average treatment effect in the treated involves transporting the risk in the untreated to the treated population. We believe that promoting a shared language, instead of using language specific for each bias, will accelerate methods development and adoption.

APPENDIX: CHAPTER 4

Appendix 4A: Rogan-Gladen Equation proof

Let $P(Y^* = 1|Y = 1) = Se$ and $P(Y^* = 0|Y = 0) = Sp$

$P(Y^* = 1)$

$= P(Y^* = 1|Y = 1)P(Y = 1) + P(Y^* = 1|Y = 0)P(Y = 0)$

$= P(Y^* = 1|Y = 1)P(Y = 1) + \big(1 - P(Y^* = 0|Y = 0)\big)\big(1 - P(Y = 1)\big)$

$= Se \times P(Y = 1) + (1 - Sp)\big(1 - P(Y = 1)\big)$

$= Se \times P(Y = 1) + 1 - P(Y = 1) - Sp + Sp \times P(Y = 1)$

$= P(Y = 1)\big(Se - (1 - Sp)\big) + (1 - Sp)$

Therefore, $P(Y = 1) = \frac{P(Y^*=1)-(1-Sp)}{Se-(1-Sp)}$

where the first equality follows the law of total probability, the second follows from convexity, the third is by definition of Se and Sp, and the fourth and fifth are algebraic rearrangement.

Conditional version

Let $P(Y^* = 1|Y = 1, S = s) = Se_S$ and $P(Y^* = 0|Y = 0, S = s) = Sp_S$

$P(Y^* = 1|S = s)$ where S is a vector of covariates

$= P(Y^* = 1|Y = 1, S = s)P(Y = 1|S = s) + P(Y^* = 1|Y = 0, S = s)P(Y = 0|S = s)$

$= P(Y^* = 1|Y = 1, S = s)P(Y = 1|S = s) + \big(1 - P(Y^* = 0|Y = 0, S = s)\big)\big(1 - P(Y = 1|S = s)\big)$

$= Se_S \times P(Y = 1|S = s) + (1 - Sp_S)\big(1 - P(Y = 1|S = s)\big)$

$= Se_S \times P(Y = 1|S = s) + 1 - P(Y = 1|S = s) - Sp_S + Sp_S \times P(Y = 1|S = s)$

$= P(Y = 1|S = s)\big(Se_S - (1 - Sp_S)\big) + (1 - Sp_S)$

Therefore, $P(Y = 1|S = s) = \frac{P(Y^*=1|S=s)-(1-Sp_S)}{Se_S-(1-Sp_S)}$

Appendix 4B: Proof of conditioning on W approach

Let $Se_{A,Z,W,R=r} = P(Y^* = 1|Y = 1, A = a, Z = z, W = w, R = r)$ and $Sp_{A,Z,W,R=r} = P(Y^* = 0|Y = 0, A = a, Z = z, W = w, R = r)$

$P(Y(a) = 1|R = 1)$

Apply law of total probability

$$= \sum_{z,w} P(Y(a) = 1|Z = z, W = w, R = 1)P(Z = z, W = w|R = 1)$$

Apply conditional exchangeability $Y(a) \coprod A|Z, W$

$$= \sum_{z,w} P(Y(a) = 1|A = a, Z = z, W = w, R = 1)P(Z = z, W = w|R = 1)$$

Apply causal consistency

$$= \sum_{z,w} P(Y = 1|A = a, Z = z, W = w, R = 1)P(Z = z, W = w|R = 1)$$

Apply Rogan-Gladen

$$= \sum_{z,w} \frac{P(Y^* = 1|A = a, Z = z, W = w, R = 1) - \left(1 - Sp_{A,Z,W,R=1}\right)}{Se_{A,Z,W,R=1} - \left(1 - Sp_{A,Z,W,R=1}\right)} P(Z = z, W = w|R = 1)$$

Apply transportability condition $R \coprod Y^*|Y, A, Z, W$

$$= \sum_{z,w} \frac{P(Y^* = 1|A = a, Z = z, W = w, R = 1) - \left(1 - Sp_{A,Z,W,R=0}\right)}{Se_{A,Z,W,R=0} - \left(1 - Sp_{A,Z,W,R=0}\right)} P(Z = z, W = w|R = 1)$$

Appendix 4C: Multinomial outcomes

Let y denote the level of a multinomial outcome, $y = 0,1, \dots, m$. To estimate $P(Y = y|A = a, Z = z, R = 1)$, we again use a maximum likelihood approach. Let $\mu_{iy} = P(Y = y|A_i, Z_i, R_i = 1)$. We specify a multinomial model consisting of logistic models for $y > 0$, $\ln(\mu_{iy}/\mu_{i0}) = \beta_{0y} + \beta_{1y}A_i + \beta_{zy}h(Z_i)$ where $h(\cdot)$ is a flexible function that may include interactions with other variables in the model. Note that $\mu_{iy} = \frac{\exp(\beta_{0y}+\beta_{1y}A_i+\beta_{zy}h(Z_i))}{[1+\sum_{k=1}^{m}\exp(\beta_{0k}+\beta_{1k}A_i+\beta_{zk}h(Z_i))]}$ when $y > 0$ and $\mu_{i0} = 1 - \sum_{k=1}^{m}\mu_{ik}$. We estimate γ by maximizing the modified likelihood in the study sample

$$L(\beta) = \prod_{i=1}^{N}\prod_{k=0}^{m}\left\{\left[\sum_{l=0}^{m}\hat{P}(Y^* = k|Y = l)\mu_{il}\right]^{R_i I(Y_i^*=k)}\right\}.$$

We estimate the risk of level y under the observed exposure distribution as $\frac{1}{n_1}\sum_{i=1}^{N}R_i\hat{\mu}_{iy}$ and the counterfactual risk of level y as $\frac{1}{n_1}\sum_{i=1}^{N}R_i\hat{\mu}_{iy}^{a}$. Note that $\sum_{k=0}^{m}\hat{\mu}_{ik} = 1$.

For estimation of the misclassification parameters in which error is differential with respect to Z, let $\rho_{iy} = P(Y^* = y|Y_i, Z_i, R_i = 0)$. In the validation data, we fit the multinomial logistic model consisting of logistic models for $y > 0$, $\ln(\rho_{iy}/\rho_{i0}) = \delta_{0y} + \sum_{k=1}^{m}\delta_{ky}I(Y_i = k) + \delta_{zy}h(Z_i)$, and transport the fitted $\hat{\delta}$ to the study sample. Each individual in the study sample will have $(m+1)^2$ misclassification parameters, one for each combination of Y^* and Y. Let $\hat{\rho}_{ikl} = \hat{P}(Y^* = l|Y_i = k, Z_i, R_i = 0)$ and

$$\hat{\rho}_{ikl} = \exp\left(\hat{\delta}_{0l} + \hat{\delta}_{kl} + \hat{\delta}_{zl}h(Z_i)\right) / \left[1 + \sum_{t=1}^{m}\exp\left(\hat{\delta}_{0t} + \hat{\delta}_{kt} + \hat{\delta}_{zt}h(Z_i)\right)\right]$$

when $l > 0$. Note for $k = 0$, there is no term $\hat{\delta}_{0l}$. For $l = 0$, $\hat{\rho}_{ik0} = 1 - \sum_{t=1}^{m}\hat{\rho}_{ikt}$.

Appendix 4D: Proof of weighted estimator

Proof 1 relying on $W \coprod Y|A,Z,R$

$P(Y^* = y|Y,A,Z,R=1)$

Apply law of total probability

$$= \sum_w P(Y^* = y|Y,A,Z,W=w,R=1)P(W=w|Y,A,Z,R=1)$$

Apply transportability condition $R \coprod Y^*|Y,A,Z,W$

$$= \sum_w P(Y^* = y|Y,A,Z,W=w,R=0)P(W=w|Y,A,Z,R=1)$$

Apply Bayes theorem

$$= \sum_w \frac{P(Y^*=y,W=w|Y,A,Z,R=0)P(Y=y,A=a,Z=z|R=0)}{P(W=w|Y,A,Z,R=0)P(Y=y,A=a,Z=z,|R=0)} P(W=w|Y,A,Z,R=1)$$

Rearrange terms

$$= \sum_w P(Y^*=y,W=w|Y,A,Z,R=0)\frac{P(W=w|Y,A,Z,R=1)}{P(W=w|Y,A,Z,R=0)}$$

Apply condition $W \coprod Y|A,Z,R$

$$= \sum_w P(Y^*=y,W=w|Y,A,Z,R=0)\frac{P(W=w|A,Z,R=1)}{P(W=w|A,Z,R=0)}$$

Apply Bayes theorem

$$= \sum_w P(Y^*=y,W=w|Y,A,Z,R=0)\frac{P(R=1|W,A,Z)P(W=w,A=a,Z=z)}{P(R=1|A,Z)P(A=a,Z=z)}\frac{P(R=0|Y,A,Z)P(A=a,Z=z)}{P(R=0|W,A,Z)P(W=w,A=a,Z=z)}$$

Rearrange terms

$$= \sum_w P(Y^*=y,W=w|Y,A,Z,R=0)\frac{P(R=1|W,A,Z)}{P(R=0|W,A,Z)}\frac{P(R=0|A,Z)}{P(R=1|A,Z)}$$

Proof 2 relying on $R \perp\!\!\!\perp Y|A,Z,W$ *and* $R \perp\!\!\!\perp Y|A,Z$

$P(Y^* = y|Y,A,Z,R=1)$

Apply law of total probability

$$= \sum_w P(Y^* = y|Y,A,Z,W=w,R=1)P(W=w|Y,A,Z,R=1)$$

Apply transportability condition $R \perp\!\!\!\perp Y^*|Y,A,Z,W$

$$= \sum_w P(Y^* = y|Y,A,Z,W=w,R=0)P(W=w|Y,A,Z,R=1)$$

Apply Bayes theorem

$$= \sum_w \frac{P(Y^* = y,W=w|Y,A,Z,R=0)P(Y=y,A=a,Z=z|R=0)}{P(W=w|Y,A,Z,R=0)P(Y=y,A=a,Z=z,|R=0)} P(W=w|Y,A,Z,R=1)$$

Rearrange terms

$$= \sum_w P(Y^* = y,W=w|Y,A,Z,R=0)\frac{P(W=w|Y,A,Z,R=1)}{P(W=w|Y,A,Z,R=0)}$$

Apply Bayes theorem

$$= \sum_w P(Y^* = y,W=w|Y,A,Z,R=0)\frac{P(R=1|Y,W,A,Z)P(Y=y,W=w,A=a,Z=z)}{P(R=1|Y,A,Z)P(Y=y,A=a,Z=z)}\frac{P(R=0|Y,A,Z)P(Y=y,A=a,Z=z)}{P(R=0|Y,W,A,Z)P(Y=y,W=w,A=a,Z=z)}$$

Rearrange terms

$$= \sum_w P(Y^* = y,W=w|Y,A,Z,R=0)\frac{P(R=1|Y,W,A,Z)}{P(R=0|Y,W,A,Z)}\frac{P(R=0|Y,A,Z)}{P(R=1|Y,A,Z)}$$

Apply condition $R \perp\!\!\!\perp Y|A,Z,W$ and $R \perp\!\!\!\perp Y|A,Z$

$$= \sum_w P(Y^* = y,W=w|Y,A,Z,R=0)\frac{P(R=1|W,A,Z)}{P(R=0|W,A,Z)}\frac{P(R=0|A,Z)}{P(R=1|A,Z)}$$

Appendix 4E: M-estimation

4E.1 M-estimation overview

Let θ denote a p-by-1-dimensional vector of parameters (i.e., $\theta = (\theta_1, \dots, \theta_p)$), O_i be the observed data where i indexes N independent individuals, and $g(\cdot)$ be a p-by-1 vector of estimating functions. An M-estimator, $\hat{\theta}$, is the solution to $\sum_{i=1}^{N} g(O_i, \theta) = 0$. For example, $g(X_i, \theta) = X_i - \theta$ is the estimating function for the mean. For maximum likelihood, the first derivative of the log-likelihood, or the score functions, are the corresponding estimating functions.

The covariance of $\hat{\theta}$, denoted as $\Sigma_{\hat{\theta}}$ (i.e., the covariance matrix for $\hat{\theta}$ where the diagonals are the variance of each parameter in $\hat{\theta}$), can be estimated by the empirical sandwich estimator $\hat{\Sigma}_{\hat{\theta}} = N^{-1}\left[B_N(\hat{\theta})^{-1} M_N(\hat{\theta}) \left\{B_N(\hat{\theta})^{-1}\right\}^T\right]$ where $B_N(\hat{\theta}) = N^{-1}\sum_{i=1}^{N} -g'(O_i; \hat{\theta})$ where $g'(O_i; \theta) = [\frac{\partial g(O_i;\theta)}{\partial \theta}]$ is the matrix of partial derivatives of g with respect to each parameter and $M_N(\hat{\theta}) = N^{-1}\sum_{i=1}^{N} g(O_i; \hat{\theta}) g(O_i; \hat{\theta})^T$. Wald-type confidence intervals for the parameters are constructed using the square root of the diagonal of $\hat{\Sigma}_{\hat{\theta}}$.

4E.2 Stacked estimating functions for estimators in the paper

Note: To simplify and limit the length of this appendix, we do not include flexible functions or all possible interactions terms. Stacks can be straightforwardly extended to accommodate flexible modeling.

4E.2.1 Notation

Let R be an indicator of the which population an individual belongs to where $R = 1$ for the study sample and $R = 0$ for the validation cohort. Here, the observed data are $O_i = \{R_i, A_i, Z_i, W_i, Y_i^*, Y_i(1 - R_i)\}$. Let $Y_i(a)$ be the potential outcome when $A_i = a$. Let $I(x)$ be an indicator function that equals 1 if x is true and 0 otherwise. For simplification of notation, Z and W are each treated as a single variable.

4E.2.2 Assuming no misclassification

Here, $\theta = (\alpha, \beta)$ with $\alpha = (\alpha_1,\ \alpha_2,\ \alpha_3, \alpha_4)$ as the parameters of interest where $\alpha_1 = \alpha_2 - \alpha_3$, $\alpha_2 = E(Y_i^1)$, $\alpha_3 = E(Y_i^0)$, and $\alpha_4 = E(Y_i)$; and $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ as the nuisance parameters for the

outcome model in the study sample, $\Pr(Y_i^* = 1) = \text{expit}(\beta_0 + \beta_1 A_i + \beta_2 Z_i + \beta_3 A_i Z_i) = \text{expit}(X_i \beta^T)$ where $X_i = (1, A_i, Z_i, A_i Z_i)$.

$$g(O_i, \theta) = \begin{bmatrix} I(R_i = 1)[(\alpha_2 - \alpha_3) - \alpha_1] \\ I(R_i = 1)(\hat{Y}(1)_i - \alpha_2) \\ I(R_i = 1)(\hat{Y}(0)_i - \alpha_3) \\ I(R_i = 1)(\hat{Y}_i - \alpha_4) \\ I(R_i = 1)[Y_i^* - \text{expit}(X_i \beta^T)] \\ I(R_i = 1)[Y_i^* - \text{expit}(X_i \beta^T)]A_i \\ I(R_i = 1)[Y_i^* - \text{expit}(X_i \beta^T)]Z_i \\ I(R_i = 1)[Y_i^* - \text{expit}(X_i \beta^T)]A_i Z_i \end{bmatrix}$$

where $\hat{Y}_i = \text{expit}\left(X_i \hat{\beta}^T\right)$, $\hat{Y}_i(0) = \text{expit}\left(X_i^0 \hat{\beta}^T\right)$, and $\hat{Y}_i(1) = \text{expit}\left(X_i^1 \hat{\beta}^T\right)$ in which $X_i^1 = (1,1, Z_i, Z_i)$ and $X_i^0 = (1,0, Z_i, 0)$.

4E.2.3 Accounting for nondifferential outcome misclassification (Figure 1A)

The full set of parameters to be estimated is $\theta = (\alpha, \beta, \gamma)$ where α and β are the same as in 4E.2.2 and $\gamma = (\gamma_1 = Se,\ \gamma_2 = 1 - Sp)$

$$g(O_i, \theta) = \begin{bmatrix} I(R_i = 1)(\alpha_2 - \alpha_3) - \alpha_1 \\ I(R_i = 1)(\hat{Y}(1)_i - \alpha_2) \\ I(R_i = 1)(\hat{Y}(0)_i - \alpha_3) \\ I(R_i = 1)\left(\hat{Y}_i - \alpha_4\right) \\ I(R_i = 1)\text{expit}(X_i \beta^T)[1 - \text{expit}(X_i \beta^T)]S_i \\ I(R_i = 1)\text{expit}(X_i \beta^T)[1 - \text{expit}(X_i \beta^T)]A_i S_i \\ I(R_i = 1)\text{expit}(X_i \beta^T)[1 - \text{expit}(X_i \beta^T)]Z_i S_i \\ I(R_i = 1)\text{expit}(X_i \beta^T)[1 - \text{expit}(X_i \beta^T)]A_i Z_i S_i \\ I(R_i = 0)I(Y_i = 1)(Y_i^* - \gamma_1) \\ I(R_i = 0)I(Y_i = 0)(Y_i^* - \gamma_2) \end{bmatrix}$$

where $S_i = \left[\frac{Y_i^*(\gamma_1 - \gamma_2)}{\gamma_2 + \text{expit}(X_i \beta^T)(\gamma_1 - \gamma_2)} - \frac{(1 - Y_i^*)(\gamma_1 - \gamma_2)}{(1 - \gamma_2) - \text{expit}(X_i \beta^T)(\gamma_1 - \gamma_2)}\right]$

Derivation of score equation for modified likelihood

The individual likelihood[101,113] is $[Se \times \mu_i + (1 - Sp) \times (1 - \mu_i)]^{R_i Y_i^*}[(1 - Se)\mu_i + Sp(1 - \mu_i)]^{R_i(1 - Y_i^*)}$ where $\mu_i = \text{expit}(X_i \beta^T)$. An individual's contribution to the log-likelihood is

$$R_i Y_i^* \log[Se \times \mu_i + (1 - Sp_i) \times (1 - \mu_i)] + R_i(1 - Y_i^*) \log[(1 - Se)\mu_i + Sp_i(1 - \mu_i)]$$

$$= R_i Y_i^* \log[(1 - Sp_i) + \mu_i(Se - (1 - Sp_i))] + R_i(1 - Y_i^*) \log[Sp_i - \mu_i(Se - (1 - Sp_i))]$$

The derivative of that log-likelihood is

$$\frac{R_i Y_i^*(Se - (1 - Sp))\frac{\partial \mu_i}{\partial \beta}}{(1 - Sp) + \frac{\partial \mu_i}{\partial \beta}(Se - (1 - Sp))} - \frac{R_i(1 - Y_i^*)(Se - (1 - Sp_i))\frac{\partial \mu_i}{\partial \beta}}{Sp_i - \frac{\partial \mu_i}{\partial \beta}(Se - (1 - Sp_i))}$$

$$= R_i \mu_i' \left[\frac{Y_i^*(Se - (1 - Sp))}{(1 - Sp) + \frac{\partial \mu_i}{\partial \beta}(Se - (1 - Sp))} - \frac{(1 - Y_i^*)(Se - (1 - Sp))}{Sp - \frac{\partial \mu_i}{\partial \beta}(Se - (1 - Sp))} \right]$$

where $\frac{\partial \mu_i}{\partial \beta_0} = \mu_i(1 - \mu_i)$; $\frac{\partial \mu_i}{\partial \beta_1} = \mu_i(1 - \mu_i)A_i$; $\frac{\partial \mu_i}{\partial \beta_2} = \mu_i(1 - \mu_i)Z_i$; $\frac{\partial \mu_i}{\partial \beta_3} = \mu_i(1 - \mu_i)A_i Z_i$

We replace Se and $1 - Sp$ with γ_1 and γ_2, respectively

$$= R_i \mu_i' \left[\frac{Y_i^*(\gamma_1 - \gamma_2)}{\gamma_2 + \mu_i(\gamma_1 - \gamma_2)} - \frac{(1 - Y_i^*)(\gamma_1 - \gamma_2)}{(1 - \gamma_2) - \mu_i(\gamma_1 - \gamma_2)} \right]$$

Let $S_i = \left[\frac{Y_i^*(\gamma_1 - \gamma_2)}{\gamma_2 + \mu_i(\gamma_1 - \gamma_2)} - \frac{(1 - Y_i^*)(\gamma_1 - \gamma_2)}{(1 - \gamma_2) - \mu_i(\gamma_1 - \gamma_2)} \right]$, then $g(X_i, \gamma) = \begin{bmatrix} R_i \mu_i(1 - \mu_i) S_i \\ R_i \mu_i(1 - \mu_i) A_i S_i \\ R_i \mu_i(1 - \mu_i) Z_i S_i \\ R_i \mu_i(1 - \mu_i) A_i Z_i S_i \end{bmatrix}$

4E.2.4 Accounting for outcome misclassification that is differential with respect to A and Z (Figure 1B)

The full set of parameters to be estimated is $\theta = (\alpha, \beta, \delta)$ where α and β are the same as in 4E.2.3 and $\delta = (\delta_0, \delta_1, \delta_2, \delta_3, \delta_4, \delta_5)$ are the nuisance parameters from the measurement error model, $\Pr(Y_i^* = 1) = \text{expit}(X_i^\dagger \delta^T)$ where $X_i^\dagger = (1, Y_i, A_i, A_i Y_i, Z_i, Z_i Y_i)$. Note the distinction that X_i is the design matrix for the outcome model and $X_i^\dagger$ is the design matrix for the measurement error model.

$$g(O_i,\theta)=\begin{bmatrix} I(R_i=1)(\alpha_2-\alpha_3)-\alpha_1 \\ I(R_i=1)\big(\hat{Y}(1)_i-\alpha_2\big) \\ I(R_i=1)\big(\hat{Y}(0)_i-\alpha_3\big) \\ I(R_i=1)\big(\hat{Y}_i-\alpha_4\big) \\ I(R_i=1)\text{expit}(X_i\beta^T)[1-\text{expit}(X_i\beta^T)]S_i \\ I(R_i=1)\text{expit}(X_i\beta^T)[1-\text{expit}(X_i\beta^T)]A_iS_i \\ I(R_i=1)\text{expit}(X_i\beta^T)[1-\text{expit}(X_i\beta^T)]Z_iS_i \\ I(R_i=1)\text{expit}(X_i\beta^T)[1-\text{expit}(X_i\beta^T)]A_iZ_iS_i \\ I(R_i=0)\big[Y_i^*-\text{expit}\big(X_i^\dagger\delta^T\big)\big] \\ I(R_i=0)\big[Y_i^*-\text{expit}\big(X_i^\dagger\delta^T\big)\big]Y_i \\ I(R_i=0)\big[Y_i^*-\text{expit}\big(X_i^\dagger\delta^T\big)\big]A_i \\ I(R_i=0)\big[Y_i^*-\text{expit}\big(X_i^\dagger\delta^T\big)\big]A_iY_i \\ I(R_i=0)\big[Y_i^*-\text{expit}\big(X_i^\dagger\delta^T\big)\big]Z_i \\ I(R_i=0)\big[Y_i^*-\text{expit}\big(X_i^\dagger\delta^T\big)\big]Z_iY_i \end{bmatrix}$$

where $S_i=\left[\frac{Y_i^*\left(\gamma_{1_i}-\gamma_{2_i}\right)}{\gamma_{2_i}+\text{expit}(X_i\beta^T)\left(\gamma_{1_i}-\gamma_{2_i}\right)}-\frac{(1-Y_i^*)(\gamma_{1_i}-\gamma_{2_i})}{\left(1-\gamma_{2_i}\right)-\text{expit}(X_i\beta^T)(\gamma_{1_i}-\gamma_{2_i})}\right]$ where $\gamma_{1_i}=\text{expit}\big(X_i^{\dagger 1}\delta^T\big)$ and $\gamma_{2i}=$ $\text{expit}\big(X_i^{\dagger 0}\delta^T\big)$ in which $X_i^{\dagger 1}=(1,1,A_i,A_i,Z_i,Z_i)$ and $X_i^{\dagger 0}=(1,0,A_i,0,Z_i,0)$.

4E.2.5 Accounting for outcome misclassification that is differential with respect to A, Z, and W (Figure 1C and 1D)

Conditioning on W

The full set of parameters to be estimated is $\theta=\{\alpha,\beta,\delta\}$ where α is the same as in 4E.2.4. $\beta=\{\beta_0,\beta_1,\beta_2,\beta_3\}$ and the design matrix for the outcome model $X_i=(1,A_i,Z_i,A_iZ_i,W_i)$. $\delta=\{\delta_0,\delta_1,\delta_2,\delta_3,\delta_4,\delta_5,\delta_6,\delta_7\}$ and the design matrix for the measurement error model $X_i^\dagger=(1,Y_i,A_i,A_iY_i,Z_i,Z_iY_i,W_i,W_iY_i)$. The stack from 4E.2.4 is expanded to include these additional parameters.

Weighted misclassification model

The full set of parameters to be estimated is $\theta=(\alpha,\beta,\delta,\nu,\phi)$ where α and β are the same as in 4E.2.4. $\delta=(\delta_0,\delta_1,\delta_2,\delta_3,\delta_4,\delta_5)$ and the design matrix for the weighted measurement error model $X_i^\dagger=(1,Y_i,A_i,A_iY_i,Z_i,Z_iY_i)$. $\nu=(\nu_0,\nu_1,\nu_2,\nu_3)$ are the nuisance parameters of the selection model where R is

the outcome with design matrix $X_i^{\ddagger} = (1, A_i, Z_i, W_i)$. $\phi = (\phi_0, \phi_1, \phi_2)$ are the nuisance parameters of the selection model for the stabilization term where R is the outcome with design matrix $X_i^{\diamond} = (1, A_i, Z_i)$.

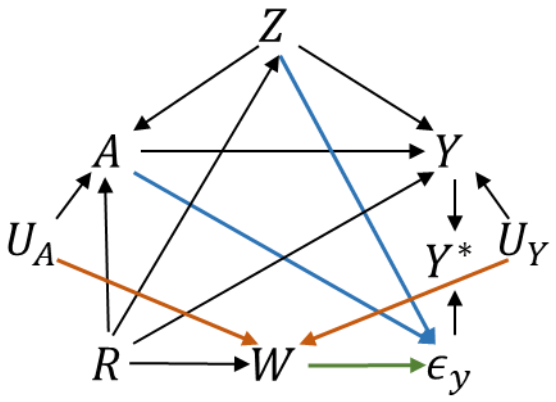
$$g(O_i, \theta) = \begin{bmatrix} I(R_i = 1)(\alpha_2 - \alpha_3) - \alpha_1 \\ I(R_i = 1)\big(\hat{Y}(1)_i - \alpha_2\big) \\ I(R_i = 1)\big(\hat{Y}(0)_i - \alpha_3\big) \\ I(R_i = 1)\big(\hat{Y}_i - \alpha_4\big) \\ I(R_i = 1)\text{expit}(X_i\beta^T)[1 - \text{expit}(X_i\beta^T)]S_i \\ I(R_i = 1)\text{expit}(X_i\beta^T)[1 - \text{expit}(X_i\beta^T)]A_iS_i \\ I(R_i = 1)\text{expit}(X_i\beta^T)[1 - \text{expit}(X_i\beta^T)]Z_iS_i \\ I(R_i = 1)\text{expit}(X_i\beta^T)[1 - \text{expit}(X_i\beta^T)]A_iZ_iS_i \\ I(R_i = 0)\big[Y_i^* - \text{expit}\big(X_i^{\dagger}\delta^T\big)\big]\pi_i \\ I(R_i = 0)\big[Y_i^* - \text{expit}\big(X_i^{\dagger}\delta^T\big)\big]\pi_iY_i \\ I(R_i = 0)\big[Y_i^* - \text{expit}\big(X_i^{\dagger}\delta^T\big)\big]\pi_iA_i \\ I(R_i = 0)\big[Y_i^* - \text{expit}\big(X_i^{\dagger}\delta^T\big)\big]\pi_iA_iY_i \\ I(R_i = 0)\big[Y_i^* - \text{expit}\big(X_i^{\dagger}\delta^T\big)\big]\pi_iZ_i \\ I(R_i = 0)\big[Y_i^* - \text{expit}\big(X_i^{\dagger}\delta^T\big)\big]\pi_iZ_iY_i \\ R_i - \text{expit}\big(X_i^{\ddagger}\nu^T\big) \\ \big[R_i - \text{expit}\big(X_i^{\ddagger}\nu^T\big)\big]A_i \\ \big[R_i - \text{expit}\big(X_i^{\ddagger}\nu^T\big)\big]Z_i \\ \big[R_i - \text{expit}\big(X_i^{\ddagger}\nu^T\big)\big]W_i \\ R_i - \text{expit}(X_i^{\diamond}\phi^T) \\ [R_i - \text{expit}(X_i^{\diamond}\phi^T)]A_i \\ [R_i - \text{expit}(X_i^{\diamond}\phi^T)]Z_i \end{bmatrix}$$

where $\pi_i = \frac{\text{expit}\left(X_i^{\ddagger}\nu^T\right)[1-\text{expit}(X_i^{\diamond}\phi^T)]}{\left[1-\text{expit}\left(X_i^{\ddagger}\nu^T\right)\right]\text{expit}(X_i^{\diamond}\phi^T)}$ and $S_i = \left[\frac{Y_i^*\left(\gamma_{1_i}-\gamma_{2_i}\right)}{\gamma_{2_i}+\text{expit}(X_i\beta^T)\left(\gamma_{1_i}-\gamma_{2_i}\right)} - \frac{(1-Y_i^*)(\gamma_{1_i}-\gamma_{2_i})}{\left(1-\gamma_{2_i}\right)-\text{expit}(X_i\beta^T)(\gamma_{1_i}-\gamma_{2_i})}\right]$,

where $\gamma_{1_i} = \text{expit}\big(X_i^{\dagger 1}\delta^T\big)$ and $\gamma_{2\,i} = \text{expit}\big(X_i^{\dagger 0}\delta^T\big)$ in which $X_i^{\dagger 1} = (1,1, A_i, A_i, Z_i, Z_i)$ and $X_i^{\dagger 0} = (1,0, A_i, 0, Z_i, 0)$.

Appendix 4F: Data generation for simulations

Appendix Figure 4F.1. Causal diagram for data generation.



Scenario A: black arrows

Scenario B: black + blue

Scenario C: black + blue + green

Scenario D: black + blue + green + orange

Generation of study sample, $R = 1$

$n_1 = 100{,}000$

$U_A \sim \text{Bernoulli}(p_{u_a} = 0.5)$

$U_Y \sim \text{Bernoulli}(p_{u_y} = 0.5)$

$Z \sim \text{Normal}(\mu_{z_1} = 0, \sigma_{z_1} = 1)$

$W \sim \text{Bernoulli}(p_{w_1})$ where $E(p_{w_1}) = 0.67$,

Scenarios A, B, C: $p_{w_1} = \text{expit}(0.693 + \ln(1)\, U_A + \ln(1)\, U_Y)$

Scenario D: $p_{w_1} = \text{expit}(0.514 + \ln(1.2)\, U_A + \ln(1.2)\, U_Y)$

$A \sim \text{Bernoulli}(p_{a_1})$ where $E(p_{a_1}) = 0.20$, $p_{a_1} = \text{expit}(-1.555 + \ln(0.6)\, Z + \ln(1.2)\, U_A)$

$Y^0 \sim \text{Bernoulli}(p_{y0_1})$ where $E(p_{y0_1}) = 0.22$, $p_{y0_1} = \text{expit}(-1.364 + \ln(1.15)\, Z + \ln(1.2)\, U_Y)$

$Y^1 \sim \text{Bernoulli}(p_{y1_1})$ where $E(p_{y1_1}) - E(p_{y0_1}) = 0.05$,

$p_{y1_1} = \text{expit}(-1.364 + \ln(1.15)\, Z + \ln(1.2)\, U_Y + \ln(0.8)\, Z + 0.275)$

$Y = AY^1 + (1 - A)Y^0$, resulting $P(Y = 1) = 0.233$

After Y^* is generated (see below), Y is set to missing in this dataset

Generation of validation data, $R = 0$

$n_0 = 2{,}000$

$U_A \sim \text{Bernoulli}(p_{u_a} = 0.5)$

$U_Y \sim \text{Bernoulli}(p_{u_y} = 0.5)$

$Z \sim N(\mu_{z_0} = 1, \sigma_{z_0} = 1)$

$W \sim \text{Bernoulli}(p_{w_0})$ where $E(p_{w_0}) = 0.33$,

Scenarios A, B, C: $p_{w_0} = \text{expit}(-0.693 + \ln(1)\, U_A + \ln(1)\, U_Y)$

Scenario D: $p_{w_0} = \text{expit}(-0.878 + \ln(1.2) U_A + \ln(1.2)\, U_Y)$

$A \sim \text{Bernoulli}(p_{x_0})$ where $E(p_{a_0}) = 0.50$, $p_{a_0} = \text{expit}(0.42 + \ln(0.6) Z + \ln(1.2)\, U_A)$

$Y^0 \sim \text{Bernoulli}(p_{y0_0})$ where $E(p_{y0_0}) = 0.11$, $p_{y0_0} = \text{expit}(-2.333 + \ln(1.15) Z + \ln(1.2)\, U_Y)$

$Y^1 \sim \text{Bernoulli}(p_{y1_0})$, $p_{y1_0} = \text{expit}(-2.333 + \ln(1.15)\, Z + \ln(1.2)\, U_Y + \ln(0.8)\, Z + 0.275)$

Resulting $P(Y^1 = 1) - P(Y^0 = 1) = 0.005$

$Y = AY^1 + (1 - A)Y^0$, resulting $P(Y = 1) = 0.115$

Generation of Y^* in both study sample and validation data

$Y^* \sim \text{Bernoulli}(p_{y^*})$

Note: first 2 parameters set so that $Se = 0.9$ and $Sp = 0.85$ marginally in validation data

Scenario A: Figure 1A

$p_{y^*} = 0.15 + 0.75Y = \text{expit}(-1.735 + 3.932Y)$

Scenario B: Figure 1B

$p_{y^*} = \text{expit}(-1.853 + 3.979Y + \ln(0.9)\, Z + \ln(1.5)\, A)$

Scenarios C: Figure 1C

$p_{y^*} = \text{expit}(-1.946 + 3.990Y + \ln(0.9)\, Z + \ln(1.5)\, A + \ln(1.3)\, W)$

Scenarios D: Figure 1D

$p_{y^*} = \text{expit}(-1.946 + 3.989Y + \ln(0.9)\, Z + \ln(1.5)\, A + \ln(1.3)\, W)$

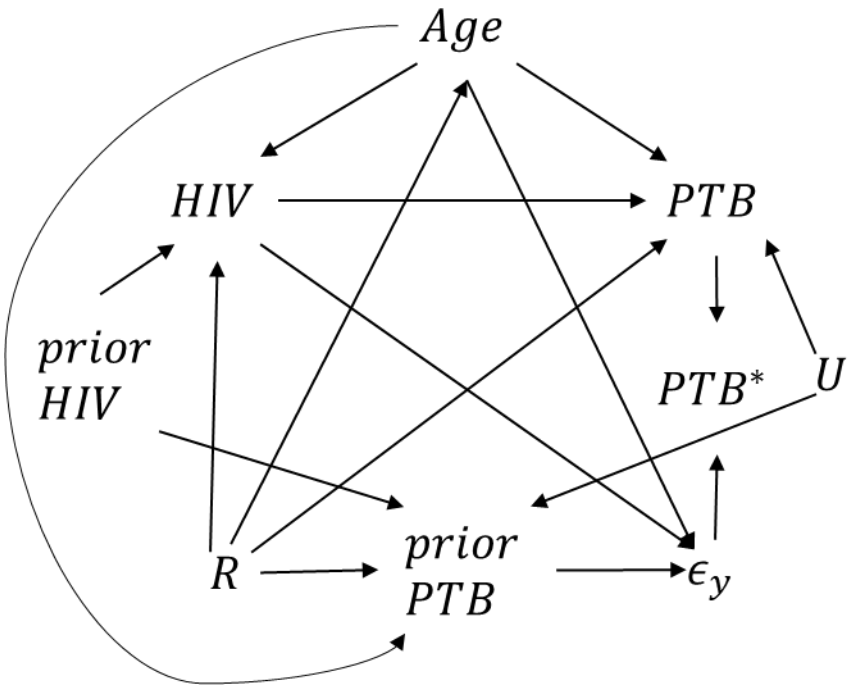
Appendix Table 4F.1. Summary of simulated study sample, $R = 1$

	Mean (min, max)			
	Se	Sp	$P(Y^* = 1)$	$E(Y^{*1} - Y^{*0})$
A	0.90	0.85	0.325	0.037
B	0.90 (0.83, 0.96)	0.85 (0.71, 0.92)	0.322	0.087
C	0.91 (0.82, 0.96)	0.84 (0.68, 0.92)	0.332	0.088
D	0.91 (0.82, 0.96)	0.84 (0.67, 0.92)	0.332	0.089

Truth: $P(Y = 1) = 0.233$ and $P(Y^1 = 1) - P(Y^0 = 1) = 0.05$

Appendix 4G: Causal diagram for applied example

Appendix Figure 4G.1. Hypothesized causal diagram for applied example.



Abbreviations: PTB, preterm birth; prior HIV, maternal HIV infection at prior pregnancy; prior PTB, preterm birth at prior pregnancy

Appendix 4H: Data generation and results for alternate Scenario D

Differences from original data generation are bolded

Generation of study sample, $R = 1$

$n_1 = 100{,}000$

$U_A \sim \text{Bernoulli}(p_{u_a} = 0.5)$

$U_Y \sim \text{Bernoulli}(p_{u_a} = 0.5)$

$Z \sim \text{Normal}(\mu_{z_1} = 0, \sigma_{z_1} = 1)$

$W \sim \text{Bernoulli}(p_{w_1})$ where $E(p_{w_1}) = 0.67$, $p_{w_1} = \text{expit}(\mathbf{-0.546 + \ln(4)\, U_A + \ln(4)\, U_Y})$

$A \sim \text{Bernoulli}(p_{a_1})$ where $E(p_{a_1}) = 0.20$, $p_{a_1} = \text{expit}(\mathbf{-2.288} + \ln(0.6)\, Z + \mathbf{\ln(4)}\, U_A)$

$Y^0 \sim \text{Bernoulli}(p_{y0_1})$ where $E(p_{y0_1}) = 0.22$, $p_{y0_1} = \text{expit}(\mathbf{-2.098} + \ln(1.15)\, Z + \mathbf{\ln(4)}\, U_Y)$

$Y^1 \sim \text{Bernoulli}(p_{y1_1})$ where $E(p_{y1_1}) - E(p_{y0_1}) = 0.05$,

$$p_{y1_1} = \text{expit}(\mathbf{-2.098} + \ln(1.15)\, Z + \mathbf{\ln(4)}\, \mathbf{U_Y} + \ln(0.8)\, Z + \mathbf{0.297})$$

Generation of validation data, $R = 0$

$n_0 = 2{,}000$

$U_A \sim \text{Bernoulli}(p_{u_a} = 0.5)$

$U_Y \sim \text{Bernoulli}(p_{u_a} = 0.5)$

$Z \sim \text{Normal}(\mu_{z_0} = 1, \sigma_{z_0} = 1)$

$W \sim \text{Bernoulli}(p_{w_0})$ where $E(p_{w_0}) = 0.33$, $p_{w_0} = \text{expit}(\mathbf{-2.227 + \ln(4) U_A + \ln(4)\, U_Y})$

$A \sim \text{Bernoulli}(p_{x_0})$ where $E(p_{a_0}) = 0.50$, $p_{a_0} = \text{expit}(\mathbf{-0.183} + \ln(0.6) Z + \mathbf{\ln(4)}\, U_A)$

$Y^0 \sim \text{Bernoulli}(p_{y0_0})$ where $E(p_{y0_0}) = 0.11$, $p_{y0_0} = \text{expit}(\mathbf{-3.112} + \ln(1.15) Z + \mathbf{\ln(4)}\, U_Y)$

$Y^1 \sim \text{Bernoulli}(p_{y1_0})$, $p_{y1_0} = \text{expit}(\mathbf{-3.112} + \ln(1.15)\, Z + \mathbf{\ln(4)}\, U_Y + \ln(0.8)\, Z + \mathbf{0.297})$

Generation of Y^* in both study sample and validation data

$Y^* \sim \text{Bernoulli}(p_{y^*})$

Note: first 2 parameters set so that $Se = 0.9$ and $Sp = 0.85$ marginally in validation data

$p_{y^*} = \text{expit}(\mathbf{-1.944} + \mathbf{3.969}Y + \ln(0.9)\, Z + \ln(1.5)\, A + \ln(1.3)\, W)$

Appendix Table 4H.1. Simulation results under the altered data generation for Scenario D (n=5000).

		Natural course					Risk difference				
Scenario[1]	Approach[2]	Mean	Bias	ESE	Avg SE	Coverage[2]	Mean	Bias	ESE	Avg SE	Coverage[2]
All	True outcome	23.3	0.0	0.1	0.1	95.3	5.0	0.0	0.4	0.4	94.4
D	Naïve analysis	33.2	9.9	0.2	0.1	0.0	9.1	4.1	0.4	0.4	0.0
	Accounting for error that is										
	Nondifferential	24.3	1.0	1.1	1.1	85.2	12.1	7.1	0.6	0.6	0.0
	Differential by A & Z	24.7	1.5	1.7	1.7	85.8	5.3	0.3	2.2	2.2	95.0
	Differential by A, Z, & W										
	Conditioning	23.2	-0.1	2.0	2.0	95.5	4.4	-0.6	2.5	2.3	94.2
	Weighted Se/Sp	23.2	-0.1	2.2	2.2	95.5	5.0	0.0	2.8	2.9	95.9

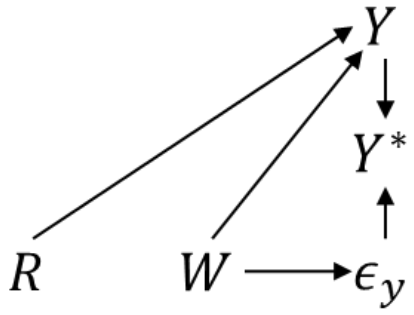
Abbreviations: ESE, empirical standard error; Avg SE, average estimated standard error; Se, sensitivity; Sp, specificity

[1]Scenarios correspond to Figure 1 Panels A, B, C, and D

[2]95% confidence interval coverage (%)

Appendix 4I: Simplified simulation

Appendix Figure 4I.1 Causal diagram of data generation for simplified simulation.



Generation of study sample, $R = 1$

$n_1 = 5{,}000{,}000$

$W \sim \text{Bernoulli}(0.5)$

$Y \sim \text{Bernoulli}(0.4 - 0.3W)$

Generation of validation data, $R = 0$

$n_0 = 5{,}000{,}000$

$W \sim \text{Bernoulli}(0.5)$

$Y \sim \text{Bernoulli}(0.8 - 0.3W)$

Generation of Y^* in both study sample and validation data

$Y^* \sim \text{Bernoulli}\big(0.7YW + 0.9Y(1-W) + 0.05(1-Y)W + 0.2(1-Y)(1-W)\big)$

Results

Truth $P(Y = 1) = 25.0$

Conditioning on W: mean estimate 25.0%, bias 0.1 percentage points

Weighted Se/Sp: mean estimate 28.0%, bias 3.0 percentage points

As expected (see proofs below), the weighted estimator is biased but conditioning on W has negligible bias.

Estimator proofs

Weighted misclassification parameters

$P(Y = 1|R = 1)$

Apply Rogan-Gladen

$$= \frac{P(Y^* = 1|R = 1) - (1 - P(Y^* = 0|Y = 0, R = 1))}{P(Y^* = 1|Y = 1, R = 1) - \big(1 - P(Y^* = 0|Y = 1, R = 1)\big)}$$

Now identify $P(Y^* = y|Y = y, R = 1)$

Apply law of total probability

$$= \sum_w P(Y^* = y|Y = y, W = w, R = 1)P(W = w|Y = y, R = 1)$$

Apply transportability condition $R \coprod Y^*|Y, W$

$$= \sum_w P(Y^* = y|Y = y, W = w, R = 0)P(W = w|Y = y, R = 1)$$

Apply Bayes theorem

$$= \sum_w \frac{P(Y^* = y, W = w|Y = y, R = 0)P(Y = y, R = 0)}{P(W = w|Y = y, R = 0)P(Y = y, R = 0)} P(W = w|Y = y, R = 1)$$

Rearrange terms, cross out common terms in numerator and denominator

$$= \sum_w P(Y^* = y, W = w|Y = y, R = 0) \frac{P(W = w|Y = y, R = 1)}{P(W = w|Y = y, R = 0)}$$

Apply $W \coprod Y|R$ (see alternative identification below)

$$= \sum_w P(Y^* = y, W = w | Y = y, R = 0) \frac{P(W = w | R = 1)}{P(W = w | R = 0)}$$

Apply Bayes theorem

$$= \sum_w P(Y^* = y, W = w | Y = y, R = 0) \frac{P(R = 1 | W = w) P(W = w)}{P(R = 1)} \frac{P(R = 0)}{P(R = 0 | W = w) P(W = w)}$$

$$= \sum_w P(Y^* = y, W = w | Y = y, R = 0) \frac{P(R = 0)}{P(R = 1)} \frac{P(R = 1 | W = w)}{P(R = 0 | W = w)}$$

Alternative identification

$$= \sum_w P(Y^* = y, W = w | Y = y, R = 0) \frac{P(W = w | Y = y, R = 1)}{P(W = w | Y = y, R = 0)}$$

Apply Bayes theorem

$$= \sum_w P(Y^* = y, W = w | Y = y, R = 0) \frac{P(R = 1 | W = w, Y = y) P(W = w, Y = y) P(R = 0 | Y = y) P(Y = y)}{P(R = 1 | Y = y) P(Y = y) P(R = 0 | W = w, Y = y) P(W = w, Y = y)}$$

$$= \sum_w P(Y^* = y, W = w | Y = y, R = 0) \frac{P(R = 0 | Y = y)}{P(R = 1 | Y = y)} \frac{P(R = 1 | W = w, Y = y)}{P(R = 0 | W = w, Y = y)}$$

Apply $R \coprod Y | W$ and $R \coprod Y$

$$= \sum_w P(Y^* = y, W = w | Y = y, R = 0) \frac{P(R = 0)}{P(R = 1)} \frac{P(R = 1 | W = w)}{P(R = 0 | W = w)}$$

Conditioning on W

$P(Y = 1|R = 1)$

Apply law of total probability

$$= \sum_{w} (Y = 1|W = w, R = 1)\, P(W = w|R = 1)$$

Apply Rogan-Gladen

$$= \sum_{w} \frac{P(Y^* = 1|W = w, R = 1) - (1 - P(Y^* = 0|Y = 0, W = w, R = 1))}{P(Y^* = 1|Y = 1, W = w, R = 1) - \big(1 - P(Y^* = 0|Y = 1, W = w, R = 1)\big)} P(W = w|R = 1)$$

Apply transportability condition $R \coprod Y^*|Y, W$

$$= \sum_{w} \frac{P(Y^* = 1|W = w, R = 1) - (1 - P(Y^* = 0|Y = 0, W = w, R = 0))}{P(Y^* = 1|Y = 1, W = w, R = 0) - \big(1 - P(Y^* = 0|Y = 1, W = w, R = 0)\big)} P(W = w|R = 1)$$

Appendix 4J: Comparison of stabilized and unstabilized weighting

Appendix Table 4J.1 Simulation results (in percentage points) comparing stabilized and unstabilized weighting under original data generation.

			Natural course					Risk difference				
Scenario[1]	Method	#	Mean	Bias	ESE	Avg SE	Coverage[2]	Mean	Bias	ESE	Avg SE	Coverage[2]
A	Stabilized	5000	23.3	0.0	2.1	2.1	95.6	5.0	0.0	2.6	2.6	95.0
	Unstabilized	4992	23.3	0.0	3.0	2.9	94.2	5.0	0.0	5.7	5.3	93.5
B	Stabilized	5000	23.3	-0.1	2.1	2.1	95.8	5.0	0.0	2.7	2.7	95.4
	Unstabilized	4992	23.3	0.0	3.2	2.8	93.4	4.8	-0.2	5.8	5.4	93.4
C	Stabilized	5000	23.3	-0.1	2.2	2.2	95.4	5.0	0.0	2.8	2.8	94.8
	Unstabilized	4989	23.3	0.0	3.3	2.9	93.3	4.8	-0.2	6.0	5.5	93.0
D	Stabilized	5000	23.3	-0.1	2.2	2.2	95.3	5.0	0.0	2.8	2.8	94.8
	Unstabilized	4987	23.3	0.0	3.0	2.9	93.0	4.8	-0.2	5.9	5.5	92.8

Abbreviations: ESE, empirical standard error; Avg SE, average estimated standard error; Se, sensitivity; Sp, specificity

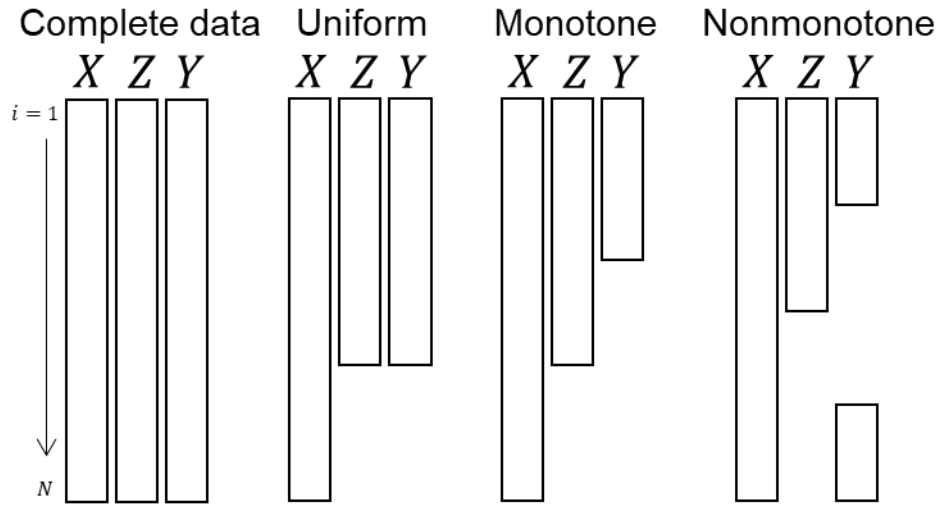
[1]Scenarios correspond to Figure 1 Panels A, B, C, and D

[2]95% confidence interval coverage (%)

APPENDIX: CHAPTER 5

Appendix 5A: Illustration of uniform, monotone, and nonmonotone missing data

Appendix Figure 5A.1. Illustration of uniform, monotone, and nonmonotone missing data.



There are three variables, X, Z, and Y. Each column represents the observed data for individuals $i = 1$ to N. When there are complete data, the columns extend the full length. When data are missing in a uniform patter, the variables will missingness (here Z and Y) are either both observed together or both missing. When data are missing in a monotone pattern, there is an ordering of the variables in which a variable is only observed if the previous variable was observed. In the illustration, Y is observed only for individuals with Z observed and Z is observed only for individuals with X observed. Uniform is a special case of monotone missingness. Finally, a nonmonotone pattern is defined as missing data that do not follow a monotone pattern.

Appendix 5B: Proof of identification of $E[Y^x]$

$$
\begin{aligned}
E[Y^x] &= E\big[E[Y^x|Z=z]\big] \\
&= E\left[\frac{\Pr(X=x|Z=z)}{\Pr(X=x|Z=z)}E[Y^x|Z=z]\right] \\
&= E\left[\frac{E[Y^x I(X=x)|Z=z]}{\Pr(X=x|Z=z)}\right] \\
&= E\left[\frac{E[YI(X=x)|Z=z]}{\Pr(X=x|Z=z)}\right] \\
&= E\left[\frac{YI(X=x)}{\Pr(X=x|Z=z)}\right],
\end{aligned}
$$

The first equality follows from the law of total probability, the second equality is multiplication by 1, the third equality follows from conditional exchangeability with positivity, the fourth equality from causal consistency, and the final equality follows from the law of total probability.

Appendix 5C: Identification of $\Pr(X = x|Z = z)$ among complete cases

$$\Pr(X = x|Z = z) = \frac{E\left[\frac{I(X = x)I(Z = z)I(R = 1)}{\Pr(R = 1|Z = z, X = x, Y = y)}\right]}{E\left[\frac{I(Z = z)I(R = 1)}{\Pr(R = 1|Z = z, X = x, Y = y)}\right]}$$

Numerator

$$E\left[\frac{I(X = x)I(Z = z)I(R = 1)}{\Pr(R = 1|Z = z, X = x, Y = y)}\right]$$
$$= E\left[E\left[\frac{I(X = x)I(Z = z)I(R = 1)}{\Pr(R = 1|Z = z, X = x, Y = y)}|Z = z, X = x, Y = y\right]\right]$$
$$= E\left[\frac{E[I(X = x)I(Z = z)I(R = 1)|Z = z, X = x, Y = y]}{\Pr(R = 1|Z = z, X = x, Y = y)}\right]$$
$$= E\left[\frac{I(X = x)I(Z = z)E[I(R = 1)|Z = z, X = x, Y = y]}{\Pr(R = 1|Z = z, X = x, Y = y)}\right]$$
$$= E\left[\frac{I(X = x)I(Z = z)\Pr(R = 1|Z = z, X = x, Y = y)}{\Pr(R = 1|Z = z, X = x, Y = y)}\right]$$
$$= E[I(X = x)I(Z = z)]$$
$$= \Pr(X = x, Z = z)$$

Denominator

$$E\left[\frac{I(Z = z)I(R = 1)}{\Pr(R = 1|Z = z, X = x, Y = y)}\right]$$
$$= E\left[E\left[\frac{I(Z = z)I(R = 1)}{\Pr(R = 1|Z = z, X = x, Y = y)}|Z = z, X = x, Y = y\right]\right]$$
$$= E\left[\frac{E[I(Z = z)I(R = 1)|Z = z, X = x, Y = y]}{\Pr(R = 1|Z = z, X = x, Y = y)}\right]$$
$$= E\left[\frac{I(Z = z)E[I(R = 1)|Z = z, X = x, Y = y]}{\Pr(R = 1|Z = z, X = x, Y = y)}\right]$$
$$= E\left[\frac{I(Z = z)\Pr(R = 1|Z = z, X = x, Y = y)}{\Pr(R = 1|Z = z, X = x, Y = y)}\right]$$
$$= E[I(Z = z)]$$
$$= \Pr(Z = z)$$

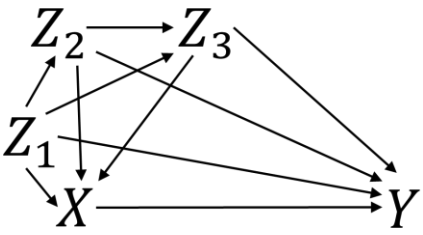
Appendix 5D: Code for simple illustrative example

https://github.com/rachael-k-ross/Wgt-NonmonotoneMiss

Appendix 5E: Details of simulation study

Generation of full data

Appendix Figure 5E.1 Causal diagram for full data generation.



Description

Let X_i be a binary random variable representing exposure (where $X = 1$ is exposed and $X = 0$ is unexposed) for individual i, $\boldsymbol{Z_i}$ represent a vector of confounders , Y_i^0 be a binary random variable representing the potential outcome when not treated ($X = 0$), Y_i^1 be a binary random variable representing the potential outcome when treated ($X = 1$), and Y_i represent the observed outcome given the observed value of X_i (i.e., $Y_i = X_i Y_i^1 + (1 - X_i)Y_i^0$). Let θ be the marginal risk difference, $E[Y^1 - Y^0]$.

We generated 5,000 trials of $N = \{1500,5000\}$. Individuals (i = 1 to N) were independent. Subscript i is suppressed for remaining description.

Let $expit(b) = 1/(1 + e^{-\mathrm{b}})$, where b is the ln $(odds)$

$\boldsymbol{Z} = \{Z_1, Z_2, Z_3\}$

$Z_1 \sim N(\mu_{z_1} = 0, sd_{z_1} = 1)$

$Z_2 \sim Bernoulli(p_{z_2})$ where $E(p_{z_2}) = 0.35$, $p_{z_2} = expit(\alpha_{20} + \alpha_{21} Z_1)$

$\alpha_{20} = -0.77$, $\alpha_{21} = \ln(0.33)$

$Z_3 \sim Bernoulli(p_{z_3})$ where $E(p_{z_3}) = 0.1$, $p_{z_3} = expit(\alpha_{30} + \alpha_{31} Z_1 + \alpha_{32} Z_2)$

$\alpha_{30} = -2.39$, $\alpha_{31} = \ln(1.6)$, $\alpha_{32} = \ln(1.4)$

$X \sim Bernoulli(p_x)$ where $E(p_x) = \{0.15, 0.50\}$, $p_x = expit(\beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3)$

For $E(p_x) = 0.15$ $\beta_0 = -2.16$

For $E(p_x) = 0.50$ $\beta_0 = -0.34$

$\beta_1 = \ln(0.9), \beta_2 = \ln(2.5), \beta_3 = \ln(1.25)$

$Y^0 \sim Bernoulli(p_{y_0})$ where $E(p_y) = 0.1$, $p_{y_0} = expit(\lambda_0 + \lambda_1 Z_1 + \lambda_2 Z_2 + \lambda_3 Z_3)$

$\lambda_0 = -2.55, \lambda_1 = \ln(0.8), \lambda_2 = \ln(1.6), \lambda_3 = \ln(2.7)$

$Y^1 \sim Bernoulli(p_{y_1})$ where $p_{y_1} = p_{y_0} + \theta$

$\theta = \{0, 0.05\}$

$Y = XY^1 + (1 - X)Y^0$

Three aspects of the data generation were varied factorially: sample size N, marginal exposure prevalence $E(p_x)$, and the exposure-outcome marginal risk difference θ

Generation of missing data

Table 5.4 in the main text shows the missing data patterns in the percent of the cohort in each pattern for the primary missing data scenario and the secondary scenarios (reproduced here).

Appendix Table 5E.1. Missing data patterns for simulation (duplicate of Table 5.4).

						% in each pattern					
Pattern (R)	X	Y	Z_1	Z_2	Z_3	Primary	Secondary scenarios				
1	O	O	O	O	O	50	65	35	50	65	35
2	M	O	O	O	O	15	10	15	10	5	15
3	O	M	O	O	O	15	10	15	10	5	10
4	M	M	O	O	O	10	5	15	10	5	10
5	O	O	O	O	M	5	5	10	5	5	10
6	M	O	O	O	M	5	5	10	5	5	10
7	O	O	M	O	O				5	5	5
8	O	M	M	O	O				5	5	5

"O" indicates variable is observed and "M" indicates variable is missing

Let R be a multinomial random variable representing which missing data pattern an individual belongs to.

Let m be the total number of patterns.

$R \sim multinomial(p_1, p_2, \ldots, p_m)$, where

$p_2 = expit(\gamma_{20} + \gamma_{21} X + \gamma_{22} Y + \gamma_{23} Z_1 + \gamma_{24} Z_2 + \gamma_{25} Z_3)$

$p_3 = expit(\gamma_{30} + \gamma_{31} X + \gamma_{32} Y + \gamma_{33} Z_1 + \gamma_{34} Z_2 + \gamma_{35} Z_3)$

$\vdots$

$p_m = expit(\gamma_{m0} + \gamma_{m1} X + \gamma_{m2} Y + \gamma_{m3} Z_1 + \gamma_{m4} Z_2 + \gamma_{m5} Z_3)$

$$p_1 = 1 - \sum_{j=2}^{m} p_j$$

For γ coefficients for each scenario, see table on the next page. The γ intercepts were set to obtain the desired marginal prevalence of each pattern. The intercept values were solved by numerical approximation using a bisection algorithm provided by Robertson et al.[135]

Note that the generation of p_2 to p_m by individual logistic regressions does not ensure structural positivity, i.e., $p_1 > 0$. We selected γ coefficients to produce bias in the complete case analysis and maintain structural positivity as possible. For the secondary scenarios with 35% complete cases, there were positivity violations when exposure prevalence was 15%; therefore, secondary scenarios were only run for exposure prevalence 50%. While we ensured structural positivity by design, there may still be random positivity violations in any simulated dataset in any scenario

Appendix Table 5E.2. γ coefficients in missing data models for simulation.

Scenario	g21	g22	g23	g24	g25	g31	g32	g33	g34	g35	g41	g42	g43	g44	g45	g51	g52	g53	g54	g55
Primary (6 patterns/50% CC)																				
MAR	0	ln(3)	0	0	0	ln(3)	0	0	0	0	0	0	0	ln(2)	ln(2)	ln(1/3)	ln(1/3)	0	0	0
MCAR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MNAR	0	ln(3)	0	0	0	ln(3)	ln(6/5)	0	0	0	ln(6/5)	ln(6/5)	0	ln(2)	ln(2)	ln(1/3)	ln(1/3)	0	0	ln(6/5)
Secondary (MAR)																				
6 patterns/65% CC	0	ln(3)	0	0	0	ln(3)	0	0	0	0	0	0	0	ln(2)	ln(2)	ln(1/3)	ln(1/3)	0	0	0
6 patterns/35% CC	0	ln(3)	0	0	0	ln(3)	0	0	0	0	0	0	0	ln(2)	ln(2)	ln(1/3)	ln(1/3)	0	0	0
8 patterns/50% CC	0	ln(3)	0	0	0	ln(3)	0	0	0	0	0	0	0	ln(2)	ln(2)	ln(1/3)	ln(1/3)	0	0	0
8 patterns/65% CC	0	ln(3)	0	0	0	ln(3)	0	0	0	0	0	0	0	ln(2)	ln(2)	ln(1/3)	ln(1/3)	0	0	0
8 patterns/35% CC	0	ln(3)	0	0	0	ln(3)	0	0	0	0	0	0	0	ln(2)	ln(2)	ln(1/3)	ln(1/3)	0	0	0

Scenario	g61	g62	g63	g64	g65	g71	g72	g73	g74	g75	g81	g82	g83	g84	g85
Primary (6 patterns/50% CC)															
MAR	0	ln(6/5)	0	ln(1/2)	0										
MCAR	0	0	0	0	0										
MNAR	ln(6/5)	ln(6/5)	0	ln(1/2)	ln(6/5)										
Secondary (MAR)															
6 patterns/65% CC	0	ln(6/5)	0	ln(1/2)	0										
6 patterns/35% CC	0	ln(6/5)	0	ln(1/2)	0										
8 patterns/50% CC	0	ln(6/5)	0	ln(1/2)	0	ln(3)	0	0	0	ln(2)	0	0	0	ln(2)	0
8 patterns/65% CC	0	ln(6/5)	0	ln(1/2)	0	ln(3)	0	0	0	ln(2)	0	0	0	ln(2)	0
8 patterns/35% CC	0	ln(6/5)	0	ln(1/2)	0	ln(3)	0	0	0	ln(2)	0	0	0	ln(2)	0

Abbreviations: CC, complete cases

Appendix 5F: Additional results tables and figures

Appendix Table 5F.1. Bias, empirical standard error, root mean squared error and 95% confidence interval coverage of the risk difference (in percentage points) for primary missing data scenario (6 patterns, 50% complete cases) when data are missing completely at random (MCAR).

	Exposure prevalence 15%					Exposure prevalence 50%				
Missing data approach[a]	Bias	ESE	RMSE	avg. ModSE	95% CI Coverage	Bias	ESE	RMSE	avg. ModSE	95% CI Coverage
Risk difference 0										
n=1500										
Full	-0.1	2.2	2.2	2.2	94%	0.0	1.6	1.6	1.6	96%
CC	-0.1	3.2	3.2	3.2	93%	0.0	2.2	2.2	2.3	95%
MI	0.1	2.9	2.9	3.0	95%	0.0	2.2	2.2	2.2	95%
Weighting[b]	-0.1	3.2	3.2	3.2	93%	0.0	2.3	2.3	2.3	95%
n=5000										
Full	0.0	1.2	1.2	1.2	95%	0.0	0.9	0.9	0.9	95%
CC	0.0	1.7	1.7	1.7	94%	0.0	1.2	1.2	1.2	95%
MI	0.0	1.6	1.6	1.6	95%	0.0	1.2	1.2	1.2	95%
Weighting[b]	0.0	1.7	1.7	1.7	94%	0.0	1.2	1.2	1.2	96%
Risk difference 0.05										
n=1500										
Full	-0.1	2.6	2.6	2.6	95%	0.0	1.7	1.7	1.8	95%
CC	0.0	3.8	3.8	3.8	94%	0.0	2.5	2.5	2.5	95%
MI	-0.2	3.4	3.4	3.5	94%	0.0	2.4	2.4	2.4	95%
Weighting[b]	-0.1	3.8	3.8	3.8	94%	0.0	2.4	2.4	2.5	95%
n=5000										
Full	0.0	1.4	1.4	1.4	95%	0.0	1.0	1.0	1.0	95%
CC	0.0	2.1	2.1	2.0	94%	0.0	1.4	1.4	1.4	95%
MI	-0.1	1.9	1.9	1.9	95%	0.1	1.3	1.3	1.3	95%
Weighting[b]	0.0	2.0	2.0	2.1	94%	0.0	1.4	1.4	1.4	95%

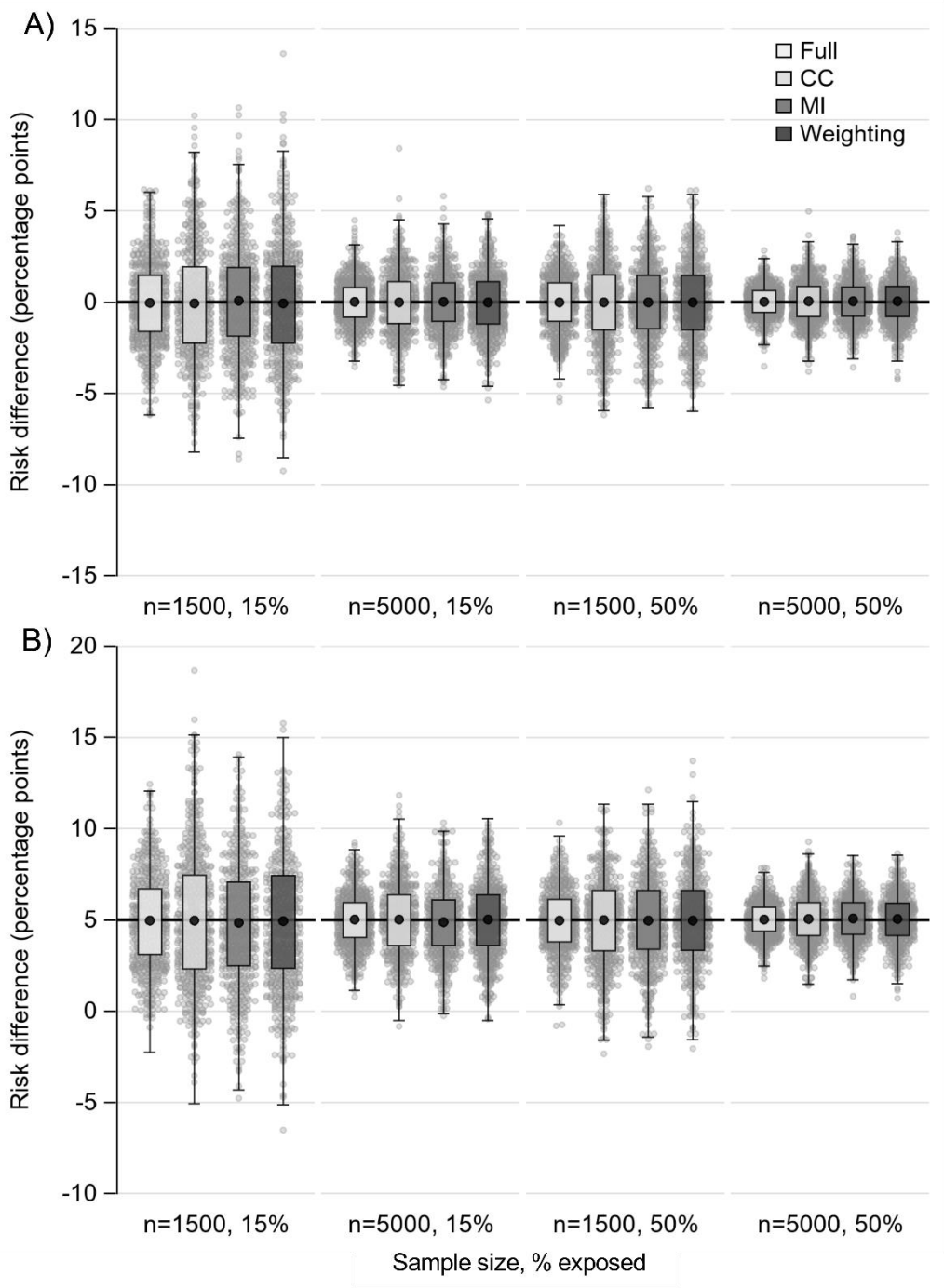
Abbreviations: CC, complete case analysis; MI, multiple imputation; ESE, empirical standard error; RMSE, root mean squared error; avg. ModSE, average model standard error; CI, confidence interval

Results from 5000 simulated cohorts

[a]All approaches addressed confounding using inverse probability of treatment weights

[b]UMLE used to estimate the missingness weights

Appendix Figure 5F.1. Boxplots of risk difference estimates for primary missing data scenario (6 patterns, 50% complete cases) when data are missing completely at random (MCAR). Panel A) true risk difference is 0%; panel B) true risk difference is 5%. Horizontal black line marks true risk difference; black dot marks mean; small gray dots are a 10% random sample of estimates. Abbreviations: CC, complete case analysis; MI, multiple imputation**.**



Appendix Table 5F.2. Bias, empirical standard error, root mean squared error and 95% confidence interval coverage of the risk difference (in percentage points) for primary missing data scenario (6 patterns, 50% complete cases) when data are missing not at random (MNAR).

	Exposure prevalence 15%					Exposure prevalence 50%				
Missing data approach[a]	Bias	ESE	RMSE	avg. ModSE	95% CI Coverage	Bias	ESE	RMSE	avg. ModSE	95% CI Coverage
Risk difference 0										
n=1500										
Full	-0.1	2.2	2.2	2.2	94%	0.0	1.6	1.6	1.6	96%
CC[b]	-2.7	2.4	3.6	2.5	67%	-1.8	1.8	2.5	1.8	83%
MI[b]	-0.9	3.8	4.0	4.3	93%	-0.9	2.6	2.7	2.6	93%
Weighting[b,c]	-2.5	4.2	4.9	4.4	80%	-1.1	2.7	2.9	2.8	94%
n=5000										
Full	0.0	1.2	1.2	1.2	95%	0.0	0.9	0.9	0.9	95%
CC	-2.8	1.3	3.1	1.3	44%	-1.7	1.0	2.0	1.0	57%
MI	-1.6	2.3	2.8	2.4	86%	-0.9	1.4	1.7	1.4	91%
Weighting[c]	-2.4	2.4	3.4	2.5	76%	-1.0	1.4	1.8	1.5	91%
Risk difference 0.05										
n=1500										
Full	-0.1	2.6	2.6	2.6	95%	0.0	1.7	1.7	1.8	95%
CC[b]	-5.5	3.1	6.3	3.1	52%	-3.7	2.0	4.2	2.0	55%
MI[b]	-2.1	4.7	5.2	4.9	88%	-1.2	2.8	3.1	2.8	93%
Weighting[b,c]	-3.5	5.3	6.3	5.4	81%	-1.4	2.9	3.2	3.1	94%
n=5000										
Full	0.0	1.4	1.4	1.4	95%	0.0	1.0	1.0	1.0	95%
CC	-5.4	1.7	5.7	1.7	15%	-3.7	1.1	3.8	1.1	10%
MI	-2.4	2.7	3.6	2.7	83%	-1.1	1.6	1.9	1.6	88%
Weighting[c]	-3.0	2.9	4.2	3.0	78%	-1.3	1.6	2.0	1.7	89%

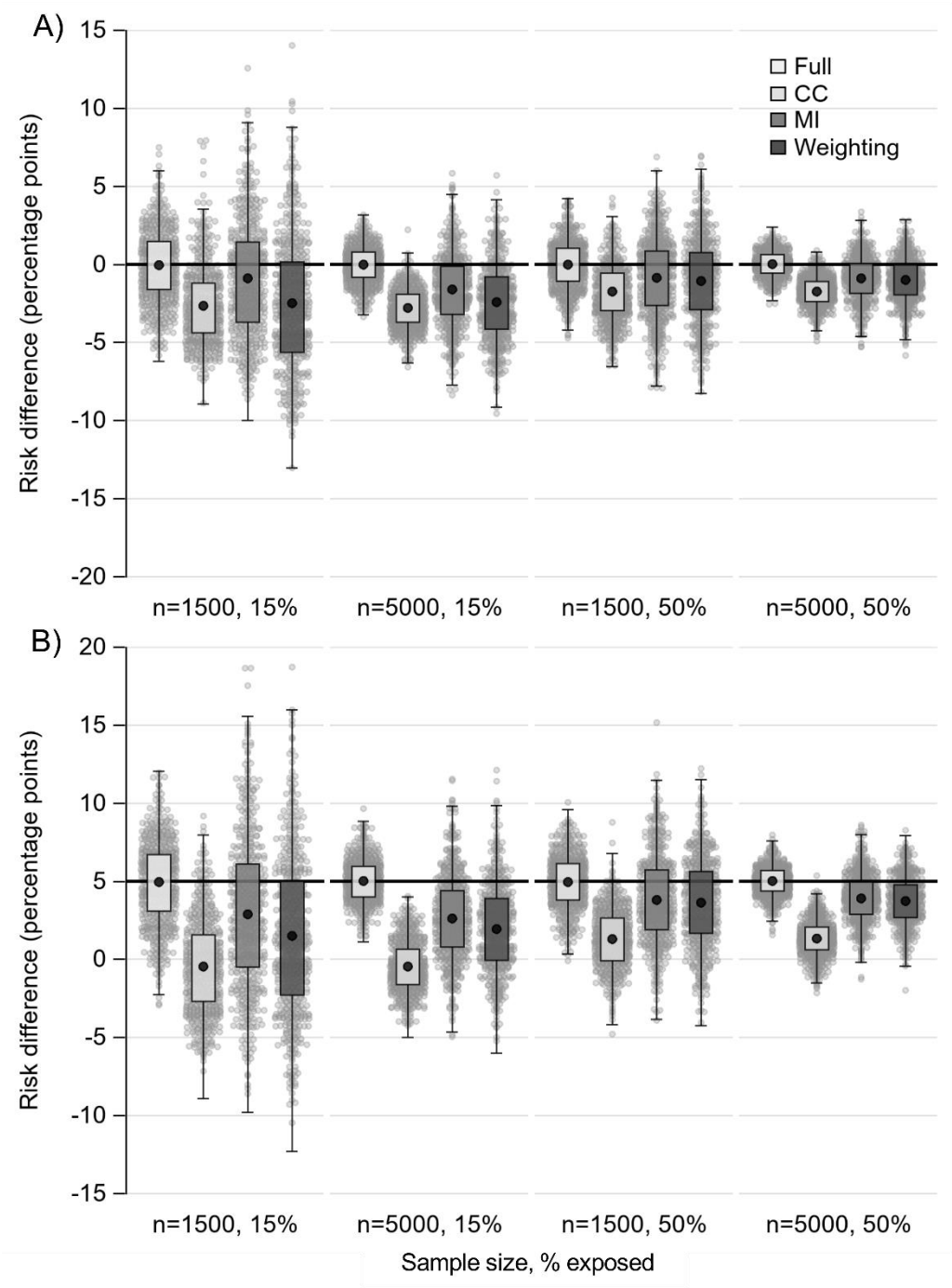
Abbreviations: CC, complete case analysis; MI, multiple imputation; ESE, empirical standard error; RMSE, root mean squared error; avg. ModSE, average model standard error; CI, confidence interval

[a]All approaches addressed confounding using inverse probability of treatment weights

Results from 5000 simulated cohorts except approaches marked with [b] at 15% prevalence (see Table 5.5 for number of failures)

[c]UMLE used to estimate the missingness weights

Appendix Figure 5F.2. Boxplots of risk difference estimates for primary missing data scenario (6 patterns, 50% complete cases) when data are missing not at random (MNAR). Panel A) true risk difference is 0%; panel B) true risk difference is 5%. Horizontal black line marks true risk difference; black dot marks mean; small gray dots are a 10% random sample of estimates. Abbreviations: CC, complete case analysis; MI, multiple imputation**.**



Appendix Table 5F.3. Bias, empirical standard error, root mean squared error and 95% confidence interval coverage of the risk difference (in percentage points) when the true risk difference is 5% and data are missing at random (MAR).

	6 patterns					8 patterns				
Missing data approach[a]	Bias	ESE	RMSE	avg. ModSE	95% CI Coverage	Bias	ESE	RMSE	avg. ModSE	95% CI Coverage
n=1500										
65% complete cases										
Full	0.0	1.7	1.7	1.8	95%	0.0	1.7	1.8	1.7	95%
CC	-1.3	2.0	2.4	2.0	91%	-0.8	2.2	2.1	2.2	94%
MI	0.1	2.3	2.3	2.3	95%	0.0	2.1	2.1	2.1	95%
Weighting[b]	0.0	2.3	2.3	2.4	96%	0.0	2.2	2.3	2.2	95%
50% complete cases										
Full	0.0	1.7	1.7	1.8	95%	0.0	1.7	1.8	1.7	95%
CC	-2.5	2.1	3.3	2.2	78%	-1.8	2.9	2.3	2.9	87%
MI	0.0	2.8	2.8	2.8	94%	0.0	2.5	2.5	2.5	95%
Weighting[b]	0.0	2.8	2.8	3.0	96%	0.0	2.7	2.8	2.7	96%
35% complete cases										
Full	0.0	1.7	1.7	1.8	95%	0.0	1.7	1.8	1.7	95%
CC	-3.8	2.4	4.5	2.4	65%	-3.3	4.2	2.5	4.2	72%
MI	0.1	3.3	3.3	3.3	94%	0.0	3.0	3.0	3.0	94%
Weighting[b]	-0.1	3.5	3.5	3.9	97%	-0.1	3.5	3.8	3.5	96%
n=5000										
65% complete cases										
Full	0.0	1.0	1.0	1.0	95%	0.0	1.0	1.0	1.0	95%
CC	-1.3	1.1	1.7	1.1	80%	-0.8	1.4	1.2	1.4	90%
MI	0.1	1.3	1.3	1.3	95%	0.1	1.2	1.2	1.2	95%
Weighting[b]	0.0	1.3	1.3	1.3	95%	0.0	1.2	1.2	1.2	95%
50% complete cases										
Full	0.0	1.0	1.0	1.0	95%	0.0	1.0	1.0	1.0	95%
CC	-2.5	1.2	2.7	1.2	45%	-1.7	2.1	1.2	2.1	72%
MI	0.1	1.5	1.5	1.5	95%	0.1	1.4	1.4	1.4	95%
Weighting[b]	0.1	1.5	1.5	1.6	96%	0.0	1.5	1.5	1.5	96%
35% complete cases										
Full	0.0	1.0	1.0	1.0	95%	0.0	1.0	1.0	1.0	95%
CC	-3.8	1.3	4.0	1.3	19%	-3.3	3.6	1.4	3.6	31%
MI	0.1	1.8	1.8	1.8	95%	0.0	1.6	1.6	1.6	94%
Weighting[b]	0.0	1.9	1.9	2.1	97%	0.0	1.9	2.0	1.9	96%

Abbreviations: CC, complete case analysis; MI, multiple imputation; ESE, empirical standard error; RMSE, root mean squared error; avg. ModSE, average model standard error; CI, confidence interval

Results from 5000 simulated cohorts

[a]All approaches addressed confounding using inverse probability of treatment weights

[b]UMLE used to estimate the missingness weights

Appendix Table 5F.4. Bias, empirical standard error, root mean squared error and 95% confidence interval coverage of the risk difference (in percentage points) when the true risk difference is 0% and data are missing at random (MAR).

	6 patterns					8 patterns				
Missing data approach[a]	Bias	ESE	RMSE	avg. ModSE	95% CI Coverage	Bias	ESE	RMSE	avg. ModSE	95% CI Coverage
n=1500										
65% complete cases										
Full	0.0	1.6	1.6	1.6	96%	0.0	1.6	1.6	1.6	96%
CC	-0.5	1.8	1.9	1.8	94%	-0.4	1.9	1.9	1.9	94%
MI	0.1	2.1	2.1	2.1	95%	0.0	1.9	2.0	1.9	95%
Weighting[b]	0.0	2.1	2.1	2.2	95%	0.0	2.0	2.1	2.0	95%
50% complete cases										
Full	0.0	1.6	1.6	1.6	96%	0.0	1.6	1.6	1.6	96%
CC	-1.1	1.9	2.2	1.9	91%	-0.8	2.2	2.0	2.2	93%
MI	0.0	2.6	2.6	2.6	94%	0.0	2.3	2.3	2.3	95%
Weighting[b]	0.0	2.7	2.7	2.7	95%	0.0	2.5	2.6	2.5	95%
35% complete cases										
Full	0.0	1.6	1.6	1.6	96%	0.0	1.6	1.6	1.6	96%
CC	-1.9	2.1	2.9	2.2	86%	-1.5	2.7	2.2	2.7	90%
MI	0.1	3.1	3.1	3.1	94%	0.1	2.8	2.8	2.8	94%
Weighting[b]	-0.2	3.4	3.4	3.6	96%	-0.1	3.3	3.5	3.3	96%
n=5000										
65% complete cases										
Full	0.0	0.9	0.9	0.9	95%	0.0	0.9	0.9	0.9	95%
CC	-0.5	1.0	1.1	1.0	92%	-0.4	1.1	1.1	1.1	94%
MI	0.0	1.2	1.2	1.2	95%	0.0	1.1	1.1	1.1	95%
Weighting[b]	0.0	1.2	1.2	1.2	96%	0.0	1.1	1.1	1.1	96%
50% complete cases										
Full	0.0	0.9	0.9	0.9	95%	0.0	0.9	0.9	0.9	95%
CC	-1.0	1.0	1.4	1.0	84%	-0.7	1.3	1.1	1.3	91%
MI	0.1	1.4	1.4	1.4	95%	0.0	1.3	1.3	1.3	95%
Weighting[b]	0.1	1.4	1.4	1.5	96%	0.0	1.4	1.4	1.4	96%
35% complete cases										
Full	0.0	0.9	0.9	0.9	95%	0.0	0.9	0.9	0.9	95%
CC	-1.9	1.2	2.2	1.2	65%	-1.5	1.9	1.2	1.9	77%
MI	0.0	1.7	1.7	1.7	94%	0.0	1.5	1.5	1.5	95%
Weighting[b]	0.0	1.8	1.8	2.0	97%	0.0	1.8	1.9	1.8	96%

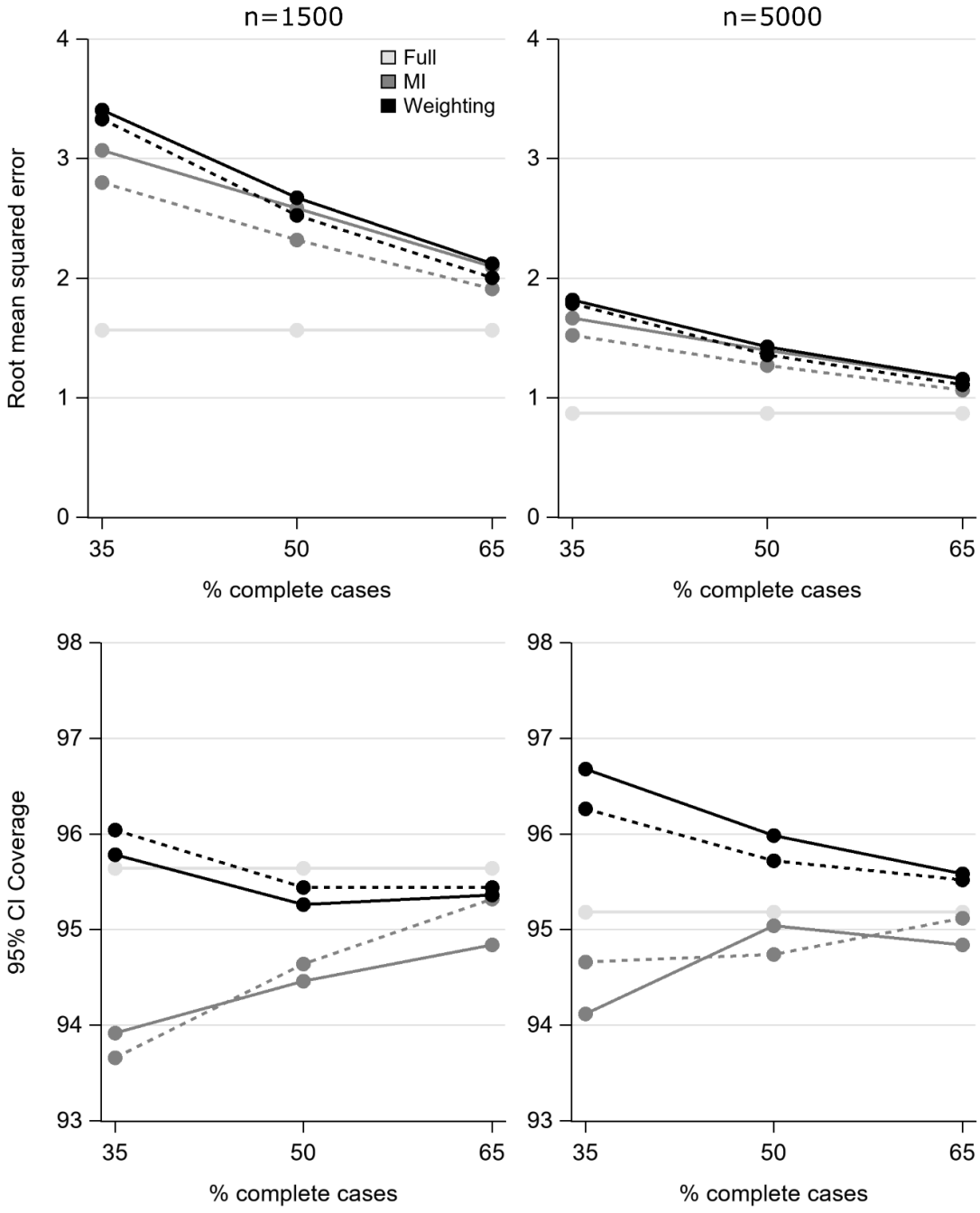
Abbreviations: CC, complete case analysis; MI, multiple imputation; ESE, empirical standard error; RMSE, root mean squared error; avg. ModSE, average model standard error; CI, confidence interval

Results from 5000 simulated cohorts except for UMLE when there were 6 patterns with 35% complete cases where 4998

[a]All approaches addressed confounding using inverse probability of treatment weights

[b]UMLE used to estimate the missingness weights

Appendix Figure 5F.3. Root mean squared error and confidence interval coverage as percent complete cases and number of patterns varies when the true risk difference is null and data are missing at random (MAR). Solid line indicates 6 patterns and dashed line indicates 8 patterns (for Full, the two lines are exactly overlaid).



Appendix Table 5F.5. Average computational time in seconds per replicate in simulation across 22 scenarios.

	R			SAS		
	Weighting[a]	MI	Relative[b]	Weighting[a]	MI	Relative[b]
Overall average	4.9	13.0	2.7	1.0	19.2	18.3
n=1500						
Average	2.7	7.4	2.8	0.5	10.3	19.3
6 patterns	2.1	7.3	3.5	0.5	10.3	22.5
8 patterns	4.3	7.8	1.8	0.7	10.2	13.8
35% complete cases	3.0	7.8	2.6	0.7	10.5	14.8
50% complete cases	2.5	7.3	3.0	0.5	9.9	20.3
65% complete cases	3.1	7.2	2.3	0.5	11.4	22.0
n=5000						
Average	7.0	18.5	2.6	1.6	28.2	18.0
6 patterns	5.8	17.3	3.0	1.3	27.4	21.7
8 patterns	10.4	21.9	2.1	2.4	30.2	12.8
35% complete cases	8.5	19.2	2.3	2.2	28.2	12.8
50% complete cases	6.6	18.4	2.8	1.3	27.4	20.4
65% complete cases	7.0	18.2	2.6	1.7	30.7	18.3

Abbreviations: MI, multiple imputation

[a]UMLE used to estimate the missingness weights

[b]Relative: MI time divided by weighting time

REFERENCES

1. Beck S, Wojdyla D, Say L, et al. The worldwide incidence of preterm birth: A systematic review of maternal mortality and morbidity. *Bull World Health Organ*. 2010;88(1):31-38. doi:10.2471/BLT.08.062554

2. Blencowe H, Cousens S, Oestergaard MZ, et al. National, regional, and worldwide estimates of preterm birth rates in the year 2010 with time trends since 1990 for selected countries: A systematic analysis and implications. *Lancet*. 2012;379(9832):2162-2172. doi:10.1016/S0140-6736(12)60820-4

3. Chawanpaiboon S, Vogel JP, Moller AB, et al. Global, regional, and national estimates of levels of preterm birth in 2014: a systematic review and modelling analysis. *Lancet Glob Heal*. 2019;7(1):e37-e46. doi:10.1016/S2214-109X(18)30451-0

4. Lee ACC, Katz J, Blencowe H, et al. National and regional estimates of term and preterm babies born small for gestational age in 138 low-income and middle-income countries in 2010. *Lancet Glob Heal*. 2013;1(1). doi:10.1016/S2214-109X(13)70006-8

5. Lawn JE, Blencowe H, Waiswa P, et al. Stillbirths: Rates, risk factors, and acceleration towards 2030. *Lancet*. 2016;387(10018):587-603. doi:10.1016/S0140-6736(15)00837-5

6. Liu L, Johnson HL, Cousens S, et al. Global, regional, and national causes of child mortality: An updated systematic analysis for 2010 with time trends since 2000. *Lancet*. 2012;379(9832):2151-2161. doi:10.1016/S0140-6736(12)60560-1

7. Hug L, Alexander M, You D, Alkema L. National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis. *Lancet Glob Heal*. 2019;7(6):e710-e720. doi:10.1016/S2214-109X(19)30163-9

8. You D, Hug L, Ejdemyr S, et al. Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: A systematic analysis by the un Inter-Agency Group for Child Mortality Estimation. *Lancet*. 2015;386(10010):2275-2286. doi:10.1016/S0140-6736(15)00120-8

9. Moster D, Lie RT, Markestad T. Long-Term Medical and Social Consequences of Preterm Birth. *N Engl J Med*. 2008;359(3):262-273.

10. Institute of Medicine. *Preterm Birth: Causes, Consequences, and Prevention*. The National Academies Press; 2007. doi:https://doi.org/10.17226/11622

11. March of Dimes, PMNCH, Save the Children, World Health Organization. *Born Too Soon: The Global Action Report on Preterm Birth*.; 2012. doi:10.2307/3965140

12. Rubens CE, Sadovsky Y, Muglia L, Gravett MG, Lackritz E, Gravett C. Prevention of preterm birth: Harnessing science to address the global epidemic. *Sci Transl Med*. 2014;6(262). doi:10.1126/scitranslmed.3009871

13. Eichenwald EC, Stark AR. Management and Outcomes of Very Low Birth Weight. *N Engl J Med*. 2008;358:1700-1711. doi:10.1097/01.aoa.0000350594.28840.06

14. Heazell AEP, Siassakos D, Blencowe H, et al. Stillbirths: Economic and psychosocial consequences. *Lancet*. 2016;387(10018):604-616. doi:10.1016/S0140-6736(15)00836-3

15. Behrman RE, Butler AS. *Preterm Birth: Causes, Consequences, and Prevention*.; 2007. doi:10.17226/11622

16. Villar J, Papageorghiou AT, Knight HE, et al. The preterm birth syndrome: A prototype phenotypic classification. *Am J Obstet Gynecol*. 2012;206(2):119-123. doi:10.1016/j.ajog.2011.10.866

17. Barros FC, Bhutta A, Batra M, Hansen TN, Victora CG, Rubens CE. Global report on preterm birth and stillbirth (3 of 7): evidence for eff ectiveness of interventions. *BMC Pregnancy Childbirth*. 2010;10(Suppl 1):1-36.

18. ClinicalTrials.gov: Improving Pregnancy Outcomes with Progesterone (IPOP) Results. Accessed July 30, 2021. https://clinicaltrials.gov/ct2/show/results/NCT03297216

19. Sreeramareddy CT, Pradhan MM, Sin S. Prevalence, distribution, and social determinants of tobacco use in 30 sub-Saharan African countries. *BMC Med*. 2014;12(1):1-13. doi:10.1186/s12916-014-0243-x

20. De Bernis L, Kinney M V., Stones W, et al. Stillbirths: Ending preventable deaths by 2030. *Lancet*. 2016;387(10019):703-716. doi:10.1016/S0140-6736(15)00954-X

21. Frederik Frøen J, Friberg IK, Lawn JE, et al. Stillbirths: Progress and unfinished business. *Lancet*. 2016;387(10018):574-586. doi:10.1016/S0140-6736(15)00818-1

22. Bhutta ZA, Das JK, Bahl R, et al. Can available interventions end preventable deaths in mothers, newborn babies, and stillbirths, and at what cost? *Lancet*. 2014;384(9940):347-370. doi:10.1016/S0140-6736(14)60792-3

23. Imdad A, Bhutta ZA. Maternal nutrition and birth outcomes: Effect of balanced protein-energy supplementation. *Paediatr Perinat Epidemiol*. 2012;26(SUPPL. 1):178-190. doi:10.1111/j.1365-3016.2012.01308.x

24. Keats EC, Haider BA, Tam E, Bhutta ZA. Multiple-micronutrient supplementation for women during pregnancy. *Cochrane Database Syst Rev*. 2019;2019(3). doi:10.1002/14651858.CD004905.pub6

25. Westreich D, Edwards JK, Lesko CR, Cole SR, Stuart EA. Target Validity and the Hierarchy of Study Designs. *Am J Epidemiol*. 2019;188(2):438-443. doi:10.1093/aje/kwy228

26. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol*. 2016;183(8):758-764. doi:10.1093/aje/kwv254

27. Hernán MA, Robins JM. *Causal Inference: What If*. Chapman & Hall/CRC; 2020.

28. DORN HF. Philosophy of inferences from retrospective studies. *Am J Public Health*. 1953;43(6.1):677-683. doi:10.2105/ajph.43.6_pt_1.677

29. Westreich D. *Epidemiology by Design: A Causal Approach to the Health Sciences*. Oxford University Press; 2020.

30. Wilcox AJ. *Fertility and Pregnancy, An Epidemiologic Perspective*. Oxford University Press; 2010.

31. Harlow SD, Ephross SA. Epidemiology of Menstruation and Its Relevance to Women ' s Health. *Epidemiol Rev*. 1995;17(2):265-286.

32. Baird DD, McConnaughey DR, Weinberg CR, et al. Application of a Method for Estimating Day of Ovulation Using Urinary Estrogen and Progesterone Metabolites. *Epidemiology*. 1995;6(5):547-550.

33. Harlow SD, Baird DD, Weinberg CR, Wilcox AJ. Urinary oestrogen patterns in long follicular phases. *Hum Reprod*. 2000;15(1):11-16.

34. Waller K, Swan SH, Windham GC, Fenster L, Elkin EP, Lasley BL. Use of Urine Biomarkers to Evaluate Menstrual Function in Healthy Premenopausal Women. *Am J Epidemiol*. 1998;147(11):1071-1080.

35. Johnson SR, Miro F, Barrett S, Ellis JE. Levels of urinary human chorionic gonadotrophin (hCG) following conception and variability of menstrual cycle length in a cohort of women attempting to conceive. *Curr Med Res Opin*. 2009;25(3):741-748. doi:10.1185/03007990902743935

36. Howards PP, Hertz-picciotto I, Weinberg CR, Poole C. Misclassification of Gestational Age in the Study of Spontaneous Abortion. *Am J Epidemiol*. 2006;164(11):1126-1136. doi:10.1093/aje/kwj327

37. Savitz DA, Terry Jr JW, Dole N, Thorp Jr JM, Siega-riz AM, Herring AH. Comparison of pregnancy dating by last menstrual period , ultrasound scanning , and their combination. *Am J Obstet Gynecol*. 2002;187(6):1660-1666. doi:10.1067/mob.2002.127601

38. Waller DK, Spears WD, Gu Y, Cunningham GC. Assessing number-specific error in the recall of onset of last menstrual period. *Paediatr Perinat Epidemiol*. 2000;14(3):263-267. doi:10.1046/j.1365-3016.2000.00275.x

39. Lynch CD, Zhang J. The research implications of the selection of a gestational age estimation method. *Paediatr Perinat Epidemiol*. 2007;21(SUPPL. 2):86-96. doi:10.1111/j.1365-3016.2007.00865.x

40. Dietz PM, England LJ, Callaghan WM, Pearl M, Wier ML, Kharrazi M. A comparison of LMP-based and ultrasound-based estimates of gestational age using linked California livebirth and prenatal screening records. *Paediatr Perinat Epidemiol*. 2007;21(SUPPL. 2):62-71. doi:10.1111/j.1365-3016.2007.00862.x

41. Ambrose CS, Caspard H, Rizzo C, Stepka EC, Keenan G. Standard methods based on last menstrual period dates misclassify and overestimate US preterm births. *J Perinatol*. 2015;35:411-414. doi:10.1038/jp.2015.25

42. Price JT, Winston J, Vwalika B, et al. Quantifying bias between reported last menstrual period and ultrasonography estimates of gestational age in Lusaka, Zambia. *Int J Gynecol Obstet*. 2019;144(1):9-15. doi:10.1002/ijgo.12686

43. Butt K, Lim KI. Guideline No. 388-Determination of Gestational Age by Ultrasound. *J Obstet Gynaecol Canada*. 2019;41(10):1497-1507. doi:10.1016/j.jogc.2019.04.010

44. Gardosi J, Geirsson RT. Routine ultrasound is the method of choice for dating pregnancy. *BJOG An Int J Obstet Gynaecol*. 1998;105:933-936. doi:10.1111/j.1471-0528.1999.tb08352.x

45. Nguyen TH, Larsen T, Engholm G, Møller H. Evaluation of ultrasound-estimated date of delivery in 17 450 spontaneous singleton births : do we need to modify Naegele ' s rule ? *Ultrasound Obstet Gynecol*. 1999;14:23-28.

46. Crowther CA, Kornman L, O'Callaghan S, George K, Furness M, Willson K. Is an ultrasound assessment of gestational age at the first antenatal visit of value? A randomised clinical trial. *Br J Obstet Gynaecol*. 1999;106:1273-1279.

47. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology*. 3rd ed. Lippincott Williams & Wilkins; 2008.

48. Van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: Five myths about measurement error in epidemiological research. *Int J Epidemiol*. 2020;49(1):338-347. doi:10.1093/ije/dyz251

49. Jurek AM, Maldonado G, Greenland S, Church TR. Exposure-measurement error is frequently ignored when interpreting epidemiologic study results. *Eur J Epidemiol*. 2006;21(12):871-876. doi:10.1007/s10654-006-9083-0

50. Brakenhoff TB, Mitroiu M, Keogh RH, Moons KGM, Groenwold RHH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. *J Clin Epidemiol*. 2018;98:89-97. doi:10.1016/j.jclinepi.2018.02.023

51. Haglund B. Birthweight distributions by gestational age: Comparison of LMP-based and ultrasound-based estimates of gestational age using data from the Swedish Birth Registry. *Paediatr Perinat Epidemiol*. 2007;21(SUPPL. 2):72-78. doi:10.1111/j.1365-3016.2007.00863.x

52. Malaba TR, Newell ML, Madlala H, Perez A, Gray C, Myer L. Methods of gestational age assessment influence the observed association between antiretroviral therapy exposure, preterm delivery, and small-for-gestational age infants: a prospective study in Cape Town, South Africa. *Ann Epidemiol*. 2018;28(12):893-900. doi:10.1016/j.annepidem.2018.08.011

53. Geerts L, Poggenpoel E, Theron G. A comparison of pregnancy dating methods commonly used in South Africa: A prospective study. *South African Med J*. 2013;103(8):552-556. doi:10.7196/SAMJ.6751

54. Gustafson P. *Measurement Error and Misclassfication in Statistics in Epidemiology: Impacts and Bayesian Adjustments*. Chapman & Hall/CRC; 2004.

55. Copeland KT, Checkoway H, Mcmichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *Am J Epidemiol*. 1977;105(5):488-495.

56. Thomas D, Stram D, Dwyer J. Exposure measurement error: Influence on exposure-disease relationships and methods of correction. *Annu Rev Public Health*. 1993;14:69-93. doi:10.1146/annurev.pu.14.050193.000441

57. Keogh RH, Shaw PA, Gustafson P, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 1—Basic theory and simple methods of adjustment. *Stat Med*. 2020;39(16):2197-2231. doi:10.1002/sim.8532

58. Brooks DR, Getz KD, Brennan AT, Pollack AZ, Fox MP. The Impact of Joint Misclassification of Exposures and Outcomes on the Results of Epidemiologic Research. *Curr Epidemiol Reports*. 2018;5(2):166-174. doi:10.1007/s40471-018-0147-y

59. Freedman LS, Midthune D, Carroll RJ, Kipnis V. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Stat Med*. 2008;27:5195-5216. doi:10.1002/sim

60. Shaw PA, Gustafson P, Carroll RJ, et al. STRATOS guidance document on measurement error and misclassification of variables in observational epidemiology: Part 2—More complex methods of adjustment and advanced topics. *Stat Med*. 2020;39(16):2232-2263. doi:10.1002/sim.8531

61. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM. *Measurement Error in Nonlinear Models : A Modern Perspective*. 2nd ed. Chapman & Hall/CRC; 2006.

62. Spiegelman D. Approaches to uncertainty in exposure assessment in environmental epidemiology. *Annu Rev Public Health*. 2010;31:149-163. doi:10.1146/annurev.publhealth.012809.103720

63. Keogh RH, White IR. A toolkit for measurement error correction, with a focus on nutritional epidemiology. *Stat Med*. 2014;33(12):2137-2155. doi:10.1002/sim.6095

64. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing Study Results: A Potential Outcomes Perspective. *Epidemiology*. 2017;28(4):553-561. doi:10.1097/ede.0000000000000664

65. Westreich D, Edwards JK, Lesko CR, Stuart EA, Cole SR. Transportability of Trial Results Using Inverse Odds of Sampling Weights. *Am J Epidemiol*. 2017;186(8):1010-1014. doi:10.1093/aje/kwx164

66. Ackerman B, Siddique J, Stuart EA. Calibrating validation samples when accounting for measurement error in intervention studies. *Stat Methods Med Res*. 2021;30(5):1235-1248. doi:10.1177/0962280220988574

67. Edwards JK, Cole SR, Shook-Sa BE, Zivich PN, Zhang N, Lesko CR. When Does Differential Outcome Misclassification Matter for Estimating Prevalence? *Epidemiology*. 2023;34(2):192-200. doi:10.1097/EDE.0000000000001572

68. Eekhout I, de Boer MR, Twisk JWR, de Vet HCW, Heymans MW. Missing Data: A Systematic Review of How They Are Reported and Handled. *Epidemiology*. 2012;23(5):729-732.

69. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: A review of current reporting and proposed guidelines. *Br J Cancer*. 2004;91(1):4-8. doi:10.1038/sj.bjc.6601907

70. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials*. 2004;1(4):368-376. doi:10.1191/1740774504cn032oa

71. Bell ML, Fiero M, Horton NJ, Hsu CH. Handling missing data in RCTs; A review of the top medical journals. *BMC Med Res Methodol*. 2014;14(1):1-8. doi:10.1186/1471-2288-14-118

72. Sullivan TR, Yelland LN, Lee KJ, Ryan P, Salter AB. Treatment of missing data in follow-up studies of randomised controlled trials: A systematic review of the literature. *Clin Trials*. 2017;14(4):387-395. doi:10.1177/1740774517703319

73. Klebanoff MA, Cole SR. Use of multiple imputation in the epidemiologic literature. *Am J Epidemiol*. 2008;168(4):355-357. doi:10.1093/aje/kwn071

74. Westreich D. Berksons bias, selection bias, and missing data. *Epidemiology*. 2012;23(1):159-164. doi:10.1097/EDE.0b013e31823b6296

75. Ross RK, Breskin A, Westreich D. When Is a Complete-Case Approach to Missing Data Valid? The Importance of Effect-Measure Modification. *Am J Epidemiol*. 2020;189(12):1583-1589. doi:10.1093/aje/kwaa124

76. Greenland S, Finkle WD. A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Analyses. *Am J Epidemiol*. 1995;142(12):1255-1264.

77. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*. 2nd ed. John Wiley & Sons, Inc.; 2002.

78. Perkins NJ, Cole SR, Harel O, et al. Principled Approaches to Missing Data in Epidemiologic Studies. *Am J Epidemiol*. 2018;187(3):568-575. doi:10.1093/aje/kwx348

79. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res*. 2013;22(3):278-295. doi:10.1177/0962280210395740

80. Harel O, Mitchell EM, Perkins NJ, et al. Multiple Imputation for Incomplete Data in Epidemiologic Studies. *Am J Epidemiol*. 2018;187(3):576-584. doi:10.1093/aje/kwx349

81. Sun BL, Perkins NJ, Cole SR, et al. Inverse-Probability-Weighted Estimation for Monotone and Nonmonotone Missing Data. *Am J Epidemiol*. 2018;187(3):585-591. doi:10.1093/aje/kwx350

82. Mansournia MA, Altman DG. Inverse probability weighting. *BMJ*. 2016;352(January):1-2. doi:10.1136/bmj.i189

83. Robins JM, Gill RD. Non-response models for the analysis of non-monotone non-ignorable missing data. *Stat Med*. 1997;16(1-3):21-37. doi:10.1002/(SICI)1097-0258(19970115)16:1<21::AID-SIM470>3.0.CO;2-F

84. Li L, Shen C, Li X, Robins JM. On weighting approaches for missing data. *Stat Methods Med Res*. 2013;22(1):14-30. doi:10.1177/0962280211403597

85. Sun BL, Tchetgen Tchetgen EJ. On Inverse Probability Weighting for Nonmonotone Missing at Random Data. *J Am Stat Assoc*. 2018;113(521):369-379. doi:10.1080/01621459.2016.1256814

86. Chi BH, Vwalika B, Killam WP, et al. Implementation of the Zambia Electronic Perinatal Record System for comprehensive prenatal and delivery care. *Int J Gynecol Obstet*. 2011;113(2):131-136. doi:10.1016/j.ijgo.2010.11.013

87. Castillo MC, Fuseini NM, Rittenhouse KJ, et al. Zambian Preterm Birth Prevention Study (ZAPPS): Cohort characteristics at enrollment. *Gates Open Res*. 2019;2(25):1-18.

88. Price JT, Vwalika B, Rittenhouse KJ, et al. Adverse birth outcomes and their clinical phenotypes in an urban Zambian cohort. *Gates Open Res*. 2020;3(1533):1-26.

89. Price JT, Vwalika B, Freeman BL, et al. Intramuscular 17-hydroxyprogesterone caproate to prevent preterm birth among HIV-infected women in Zambia: Study protocol of the IPOP randomized trial. *BMC Pregnancy Childbirth*. 2019;19(1):1-9. doi:10.1186/s12884-019-2224-8

90. Price JT, Vwalika B, Freeman BL, et al. Weekly 17 alpha-hydroxyprogesterone caproate to prevent preterm birth among women living with HIV: a randomised, double-blind, placebo-controlled trial. *Lancet HIV*. 2021;8(10):e605-e613. doi:10.1016/S2352-3018(21)00150-8

91. Papageorghiou AT, Kennedy SH, Salomon LJ, et al. International standards for early fetal size and pregnancy dating based on ultrasound measurement of crown-rump length in the first trimester of pregnancy. *Ultrasound Obstet Gynecol*. 2014;44(6):641-648. doi:10.1002/uog.13448

92. Papageorghiou AT, Kemp B, Stones W, et al. Ultrasound-based gestational-age estimation in late pregnancy. *Ultrasound Obstet Gynecol*. 2016;48(6):719-726. doi:10.1002/uog.15894

93. Hadlock FP, Deter RL, Harrist RB, Park SK. Estimating fetal age: Computer-assisted analysis of multiple fetal growth parameters. *Radiology*. 1984;152(2):497-501. doi:10.1148/radiology.152.2.6739822

94. Hadlock FP, Shah YP, Kanon DJ, Lindsey J V. Fetal Crown-Rump Length: Reevaluation of Relation to Menstural Age (5-18 weeks) with High-Resolution Real-Time US. *Radiology*. 1992;182:501-505.

95. Scholl TO, Hediger ML. Anemia and iron-deficiency anemia: Compilation of data on pregnancy outcome. *Am J Clin Nutr*. 1994;59(suppl):492-501. doi:10.1093/ajcn/59.2.492S

96. Klebanoff MA, Shiono PH, Selby J V., Trachtenberg AI, Graubard BI. Anemia and spontaneous preterm birth. *Am J Obstet Gynecol*. 1991;164:59-63. doi:10.1016/0002-9378(91)90626-3

97. Goldenberg RL, Rouse DJ. Prevention of Premature Birth. *N Engl J Med*. 1998;338(5):313-320.

98. Klebanoff MA, Shiono PH, Berendes HW, Rhoads GG. Facts and Artifacts About Anemia and Preterm Delivery. *JAMA*. 1989;262(4):511-515. doi:10.1001/jama.1990.03440090030018

99. The American College of Obstetricians and Gynecologists. *Anemia in Pregnancy*. Vol 138.; 2021.

100. Michels KB. A renaissance for measurement error. *Int J Epidemiol*. 2001;30(3):421-422. doi:10.1093/ije/30.3.421

101. Lyles RH, Tang L, Superak HM, et al. Validation data-based adjustments for outcome misclassification in logistic regression: An illustration. *Epidemiology*. 2011;22(4):589-598. doi:10.1097/EDE.0b013e3182117c85

102. Lyles RH, Zhang F, Drews-Botsch C. Combining internal and external validation data to correct for exposure misclassification: A case study. *Epidemiology*. 2007;18(3):321-328. doi:10.1097/01.ede.0000260004.49431.70

103. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health*. 2016;37:61-81. doi:10.1146/annurev-publhealth-032315-021353

104. Young JC, Conover MM, Jonsson Funk M. Measurement Error and Misclassification in Electronic Medical Records: Methods to Mitigate Bias. *Curr Epidemiol Reports*. 2018;5(4):343-356. doi:10.1007/s40471-018-0164-x

105. Rudolph JE, Cartus A, Bodnar LM, Schisterman EF, Naimi AI. The Role of the Natural Course in Causal Analysis. *Am J Epidemiol*. 2022;191(2):341-348. doi:10.1093/aje/kwab248

106. Stringer EM, Chibwesha C, Stoner MCD, et al. A population-based cohort study of stillbirth among twins in Lusaka, Zambia. *Int J Gynecol Obstet*. 2015;130(1):74-78. doi:10.1016/j.ijgo.2014.12.015

107. Westreich D, Cole SR. Invited commentary: Positivity in practice. *Am J Epidemiol*. 2010;171(6):674-677. doi:10.1093/aje/kwp436

108. Cole SR, Frangakis CE. The consistency statement in causal inference: A definition or an assumption? *Epidemiology*. 2009;20(1):3-5. doi:10.1097/EDE.0b013e31818ef366

109. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. *Am J Epidemiol*. 1978;107(1):71-76.

110. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol*. 1996;25(6):1107-1116. doi:10.1093/ije/25.6.1107

111. Greenland S. Quantifying biases in causal models: classical confounding. *Epidemiology*. 2003;14(3):300-306.

112. Lyles RH, Lin J. Sensitivity analysis for misclassification in logistic regression via likelihood methods and predictive value weighting. *Stat Med*. 2010;29(22):2297-2309. doi:10.1002/sim.3971

113. Edwards JK, Cole SR, Fox MP. Flexibly accounting for exposure misclassification with external validation data. *Am J Epidemiol*. Published online 2020.

114. Snowden JM, Rose S, Mortimer KM. Implementation of G-Computation on a Simulated Data Set: Demonstration of a Causal Inference Technique. *Am J Epidemiol*. 2011;173(7):731-738. doi:10.1093/aje/kwq472

115. Stefanski LA, Boos DD. The Calculus of M-Estimation. *Am Stat*. 2002;56(1):29-38.

116. Cole SR, Edwards JK, Breskin A, et al. Illustration of Two Fusion Designs and Estimators. *Am J Epidemiol*. Published online 2022:3. doi:10.1093/aje/kwac067

117. Morris TP, White IR, Crowther MJ. Using simulation studies to evaluate statistical methods. *Stat Med*. 2019;38(11):2074-2102. doi:10.1002/sim.8086

118. Howe CJ, Cole SR, Westreich D, Greenland S, Napravnik S, Eron JJ. Splines for Trend Analysis and Continuous Confounder Control. *Epidemiology*. 2011;22(6):874-875. doi:10.1097/EDE.0b013e31823029dd

119. Mitra N, Roy J, Small D. The Future of Causal Inference. *Am J Epidemiol*. 2022;191(10):1671-1676. doi:10.1093/aje/kwac108

120. Degtiar I, Rose S. A Review of Generalizability and Transportability. *Annu Rev Stat Its Appl*. 2023;10(1):1-30. doi:10.1146/annurev-statistics-042522-103837

121. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology*. 2003;14(6):680-686. doi:10.1097/01.EDE.0000081989.82616.7d

122. Price JT, Vwalika B, Edwards JK, et al. Maternal HIV Infection and Spontaneous Versus Provider-Initiated Preterm Birth in an Urban Zambian Cohort. *J Acquir Immune Defic Syndr*. 2021;87(2):860-868. doi:10.1097/QAI.0000000000002654

123. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60:578-586. doi:10.1136/jech.2004.029496

124. Ross RK, Breskin A, Breger TL, Westreich D. Reflection on modern methods: combining weights for confounding and missing data. *Int J Epidemiol*. Published online 2021:1-6. doi:10.1093/ije/dyab205

125. R Core Team. R: A language and environment for statistical computing. Published online 2022. http://www.r-project.org/

126. Gelfand AE, Smith AFM, Lee T-M. Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling. *J Am Stat Assoc*. 1992;87(418):523. doi:10.2307/2290286

127. Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique and future directions. *Stat Med*. 2009;28:3049-3067. doi:10.1002/sim

128. Su Y-S, Yajima M. R2jags. Published online 2021. https://cran.r-project.org/package=R2jags

129. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling. In: *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*. ; 2003. https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

130. Chib S, Greenberg E. Understanding the Metropolis-Hastings Algorithm. *Am Stat*. 1995;49(4):327-335.

131. Austin PC. Variance estimation when using inverse probability of treatment weighting (IPTW) with survival analysis. *Stat Med*. 2016;35(30):5642-5655. doi:10.1002/sim.7084

132. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc*. 1994;89(427):846-866. doi:10.1080/01621459.1994.10476818

133. Efron B, Tibshirani R. Boostrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Stat Sci*. 1986;1(1):54-77. https://projecteuclid.org/download/pdf_1/euclid.ss/1177013437

134. Johnson RW. An Introduction to the Bootstrap. *Teach Stat*. 2001;23(2):49-54. doi:10.1111/1467-9639.00050

135. Robertson SE, Dahabreh IJ, Steingrimsson JA. Using numerical methods to design simulations: revisiting the balancing intercept. *Am J Epidemiol*. Published online 2021. https://pubmed.ncbi.nlm.nih.gov/28459981/

136. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res*. 2011;20(1):40-49. doi:10.1002/mpr

137. Raghunathan TE, Lepkowski JM, Hoewyk J Van, Solenberger P. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Surv Methodol*. 2001;27(1):85-95.

138. van Buuren S. mice: Multivariate Imputation by Chained Equations. Published online 2021. https://cran.r-project.org/package=mice

139. Rubin DB, Schenker N. Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *J Am Stat Assoc*. 1986;81(394):366-374. doi:10.1080/01621459.1986.10478280

140. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Stat Sci*. 1992;7(4):457-511.

141. Cole SR, Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008;168(6):656-664. doi:10.1093/aje/kwn164

142. Allison PD. Handling Missing Data by Maximum Likelihood. *SAS Glob Forum 2012 Stat Data Anal*. Published online 2012:1-21.

143. Bernaards CA, Belin TR, Schafer JL. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat Med*. 2007;26:1368-1382. doi:10.1002/sim

144. Stuart EA, Azur MJ, Frangakis C, Leaf P. Multiple imputation with large data sets: A case study of the children's mental health initiative. *Am J Epidemiol*. 2009;169(9):1133-1139. doi:10.1093/aje/kwp026

145. Von Hippel PT. How to impute interactions, squares, and other transformed variables. *Sociol Methodol*. 2009;39(1):265-291. doi:10.1111/j.1467-9531.2009.01215.x

146. Seaman SR, Bartlett JW, White IR. Multiple imputation of missing covariates with non-linear effects and interactions: An evaluation of statistical methods. *BMC Med Res Methodol*. 2012;12(Mi):1-13. doi:10.1186/1471-2288-12-46

147. White IR, Royston P, Wood AM. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*. 2011;30(4):377-399. doi:10.1002/sim.4067

148. Allison PD. Imputation of categorical variables with PROC MI. In: *30th Meeting of SAS Users Group International*. Vol 113. ; 2005:1-14.

149. Lee KJ, Carlin JB. Multiple imputation for missing data: Fully conditional specification versus multivariate normal imputation. *Am J Epidemiol*. 2010;171(5):624-632. doi:10.1093/aje/kwp425

150. Horton NJ, Lipsitz SR, Parzen M. A potential for bias when rounding in multiple imputation. *Am Stat*. 2003;57(4):229-232. doi:10.1198/0003130032314

151. Graham JW. Missing data analysis: Making it work in the real world. *Annu Rev Psychol*. 2009;60:549-576. doi:10.1146/annurev.psych.58.110405.085530

152. Rubin DB. Randomization analysis of experimental data: The fisher randomization test. *J Am Stat Assoc*. 1980;75(371):575-582. doi:10.1080/01621459.1980.10477512