Bias bounds and target trials for causal inference in observational epidemiology

A dissertation presented

by

Louisa Hills Smith

to

The Department of Epidemiology

in partial fulfillment of the requirements

for the degree of

PhD

in the subject of

Population Health Sciences

Harvard University

Cambridge, Massachusetts

May 2021

Dissertation Advisors:
Miguel Hernán
Tyler J. VanderWeele

Author:
Louisa Hills Smith

Bias bounds and target trials for causal inference in observational epidemiology

Abstract

Observational epidemiology is critical for understanding population health but requires careful consideration of possible biases. Tools for avoiding and managing these biases are essential. This dissertation describes and implements methods for designing, analyzing, and assessing observational studies, with a particular focus on target-trial emulation and bounds for biases.

In Chapter 1, I investigate the association between COVID-19 and preterm birth using data from a large, international pregnancy registry. The principles of target-trial emulation guide the design of an analysis that avoids immortal time bias while allowing for the evaluation of gestational age-specific effects of the disease. I show that severe COVID-19 in the third trimester increases risk of preterm birth, but carries less risk earlier in pregnancy, and mild or moderate COVID-19 confers minimal added risk at any time during pregnancy. This conclusion is confirmed with additional, complementary analyses.

Chapter 2 concerns a more complex target trial that implements sustained treatment strategies. In the setting of recurrent prostate cancer, I design a trial to estimate the optimal approach for initiating hormonal treatment based on biomarker characteristics. I then describe and conduct its emulation using two complementary methods: the parametric g-formula and inverse probability-weighted dynamic marginal structural models. I find no evidence that any of the treatment strategies I consider improves upon the approach of initiating treatment only with evidence of overt metastasis.

Finally, in Chapter 3, I improve upon existing methods for sensitivity analysis that can be used to assess one type of bias at a time. Building on the E-value approach for unmeasured confounding, as well as similar bounds for selection bias and misclassification, I consider the effects of these three biases simultaneously. I show that a bound for the bias of the observed risk ratio can be constructed as a function of sensitivity analysis parameters describing each type of bias. I apply this method for sensitivity analysis to studies of exposures in pregnancy and demonstrate the software developed to implement it.

Contents

List of Tables

List of Figures

Acknowledgments

When I abruptly found out that my doctoral experience would look nothing like I had been expecting, when I learned that I'd be starting intense cancer treatment midway through my second year, when I calculated that I'd be taking my qualifying exams halfway through five months of chemotherapy, my doctor told me to stop worrying about all that: "Our goal is to make sure you get a chance to actually *use* your PhD," she said. Since then my focus has been on making it here, to the time when I can write these words and acknowledge the people who have given me the support, tools, and opportunities that allowed me to reach this point despite the obstacles.

First, my dissertation committee: Miguel Hernán, who was teaching me through his writing well before I met him. In the years since, he has taught me to be a stronger, clearer writer, teacher, and thinker – and how to have fun after a symposium. Sonia Hernández-Díaz, whose enthusiasm and focus on answering important questions with the necessary rigor has made the past year so thrilling. And Tyler VanderWeele, who never doubted that I would succeed even when it felt impossible, who always knew what to say to lift me up, but whose high expectations and insightful questions made me a better researcher – and person.

I've been lucky to have had fantastic teachers my whole life. Some of them I've not only learned from as a student but have had the privilege of teaching alongside. Jamie Robins has taught me to appreciate the subtleties and be precise with my language and notation. He made sure that I was supported and recognized for my teaching work in his course. Jarvis Chen, along with the rest of the PHS 2000 team, has poured endless amounts of energy into his teaching. I am grateful to have been able to watch him develop the course and learn in the process. I've never felt like more of an imposter than when I first stepped into Andrea Rotnitzky's class and watched

in awe as she covered the whiteboards with her brilliance. I am so grateful she trusted me to learn from her and eventually to help her teach.

I had the pleasure of collaborating with a number of other epidemiologists, biostatisticians, clinicians, and more over the past few years. With respect to the papers in this dissertation, I'd like to thank Camille Dollinger, Diego Wyszynski, Xabier García de Albéniz, and Maya Mathur for patiently answering questions, helping me dig through data, and making this work more fun.

When it comes to making work fun, I could not have asked for better than my epi cohort. My co-magistrate Susanna Mitro knows exactly how to make things more fun (more rules!) and is responsible for so much laughter (probably too much, if you ask the people whose offices shared walls with the epi student room). I had to desert my quals study group early on, but Joy Shi, Jenny Sun, and Christine Tedjianto taught me everything I needed to know in just three sessions right before the exam. Beyond their brilliance, they are also a delight to hang out with. From those first days taking over a row of Kresge G2 in PHS 2000 to our Zoom hangouts this past year, I've been lucky to have such an incredible cohort, which also includes Emma Accorsi, Geetha Iyer, Tiffany Lemon, and Tian-Shin (Cindy) Yeh (and of course Hillary Ditmars!).

In the epi department, Eric DiGiovanni and Caroline Huntington, and earlier Ellen Furxhi, made the 9th floor a lively but administratively efficient home base for me, where I never lacked anything I needed, whether batteries or beer. Thank you to the classmates who made the epi student room a place where I could get my hardest epi questions answered, learn the latest gossip, nap on a pilfered beanbag, get advice about difficult conversations or emails, and always laugh. Sam Molsberry, Kelsey Vercammen, and Jack Cordes were other devoted leaders of the John Graunt Society whose dedication to good times and a sense of community I'm grateful for. Dale Barnhart, Amanda Markovitz, Ruoran Li, and Chris Boyer helped lead the Epi Methods Journal Club through many thought-provoking conversations (and games of Pandemic).

I am thankful to have been part of the inaugural Population Health Sciences cohort, which allowed me to meet so many scholars and make so many friends in other departments — too many

to name, but a particular thanks to Gabe Schwartz and Emma Clarke, great teachers and friends with whom conversations have been stimulating. I'm so glad I got to live with MK Quinn! Bruce Villineau and Matthew Boccuzzi have kept things running smoothly for the PHS program despite all the challenges.

Thanks go as well to the biostatistics department, including Jelena Follweiler for friendly administrative guidance, Lee Kennedy-Shaffer and Dustin Rabideau for the lounge lunches, and all of the other classmates and TAs whose help saved me in homeworks and labs. Sebastien Haneuse was patient and kind when I enrolled in and then dropped out of his class, after which he still agreed to be on my oral exam committee.

Before Harvard, I had the best preparation possible at UC Berkeley. I wouldn't be anywhere if Barbara Abrams had not taken a chance on me and then given me so much freedom to learn what epidemiology had to offer. I am particularly grateful she put me in an office (and sometimes a closet) with Stephanie Leonard and Lucia Petito, who started off as a source of seemingly endless epi and biostats knowledge and soon became some of my closest friends. I'm also grateful to Holly Elser for always making me sit in the front row, to the other students who welcomed me into their fold and to their happy hours, and to the other teachers and mentors at Berkeley who introduced me to so many new ideas.

Outside of epidemiology, I am lucky to draw support from so many great friends. Hanging out with Griffin Gorsky and Grace Hyndman feels like home no matter where we are, their families like an extension of mine. Kyle Love persists and always will through the friendship of Kate Kelly, Megan Martinez, and (again!) Grace Hyndman. I'm so grateful for adventures with Jenny Desrosier, Emily Lamb, and Chelsea Macco, and I can't wait for the next one. I don't know what I would have done without Kate Fritzsche's understanding over the past few years; she is an incredible example of both strength and vulnerability for me and so many others. And to all my other friends out there — thank you for the postcards, the texts, the baby and puppy photos, the

calls, the visits, the meals, the socks, everything that has helped get me here; know that I am truly grateful and wish I could name you all.

The students and families of Little Wound School instilled in me a love of teaching and sent me on this journey. *Mitakuye oyasin.*

My grandparents, my many aunts and uncles, and my other relatives and the friends who might as well be, have done everything from helping fund my education, to feeding me Sunday dinners, to letting me live in a spare room, to sending me care packages — the list goes on and on. My brother, Peter Smith, cooks for me whenever he comes home and takes me on adventures whenever I visit — I hope that can happen again soon. As for my parents, Megan Thorn and Ned Smith, I never would have expected to have spent this much time with them during my PhD program, but how grateful am I to have been able to. They are generous, caring, encouraging, and fun to be around; I am so lucky.

Finally, who else allowed me to reach this day, when I can put my PhD to use? It was the doctors who have treated me like a full person: Dr. Houlihan, Dr. Patel, Dr. Recht, Dr. Singhal. It was the nurses, nurse practitioners, physician assistants, nursing assistants, techs who have devoted their lives to caring for me and others, particularly Eileen, Jill, Ryan, Rebecca, Zinat, Tilly, and everyone else who made me actually look forward to visiting that other 9th floor. It was the scientists who discovered the drugs I've taken and the clinicians who conducted the trials that determined my treatments. And, most of all, it was the participants in those trials who bravely stepped into the unknown for the chance at someday saving the lives of people like me. This dissertation is dedicated to them, to all the participants in research who've made my career possible in more ways than one.

Introduction

Most evidence about what improves and impairs human health is produced from observational data. From electronic health records to internet surveys, insurance claims to smartphone tracking, data that can be used to answer questions about what makes us healthy is everywhere. One difficulty when using these data to address such questions is that exposures affecting health — diet, medications, habits, toxins — are not evenly distributed across people, place, and time. Unlike the selection of drug or placebo in a clinical trial, such exposures are not randomized within a well-defined group of eligible participants. An important element of observational epidemiology is the development and use of tools to understand cause-and-effect relationships in health when we can't run a randomized controlled trial.

Two tools in particular can help us better use observational data for causal inference. First, target trials can help us design better observational studies by forcing us to imagine how we would design and analyze a randomized controlled trial to answer the same question. Second, using bounds for sensitivity analysis helps show us how robust our results are to possible biases. This dissertation applies and advances these tools.

In **Chapter 1** I investigate the relationship between COVID-19 and preterm birth using data from a large, internet-based pregnancy cohort. Reports early in the pandemic suggested that people with COVID-19 in pregnancy were more likely than expected to deliver preterm. These initial case reports were descriptive in nature, describing how many patients had been admitted

with COVID-19 and given birth and, of those, how many were preterm. Descriptive research of this sort is critical as we seek to understand a novel disease. However, when it comes to interpreting the statistics generated by these studies or making causal claims about the effects of COVID-19, careful attention is required.

Consider what we might see in the data if COVID-19 did not affect pregnancy duration. Suppose instead that time of delivery was decided at conception and did not change, whether or not a pregnant individual developed COVID-19. If COVID-19 randomly spread through the population, it would affect with equal chance those who were destined to give birth at 32 weeks as those who would carry their pregnancies until 42 weeks. However, the former are more likely to have given birth before any COVID-19 exposure reached them, while the latter have an additional ten weeks to be infected by the virus. If we compared the number of preterm and term deliveries that had been exposed to COVID-19, we would therefore find that relatively more term deliveries had been exposed. Because COVID-19 is less likely to have occurred in preterm deliveries, it would appear to have *benefits* for pregnancy duration.

Other patterns of exposure and delivery are of course possible and could result in similarly misleading conclusions. For example, a study of people with COVID-19 at delivery might find that preterm birth is relatively uncommon, even with severe disease, if most of the sample had already passed the 37-week threshold that defines preterm delivery when exposed. By contrast, a study of pregnant patients hospitalized with COVID-19 might find the risk of preterm delivery to be quite high if only those who give birth during the same hospitalization episode are included and those who are discharged with ongoing pregnancies excluded because preterm delivery cannot yet be assessed.

If, however, we imagine how we might design a trial to investigate this question, the proper choice of sample and analysis is clarified. Designing such a trial requires that we step away from reality and imagine that we have the capability (and ethics approval) to infect pregnant individuals with SARS-CoV-2, the virus that causes COVID-19, and even force them to develop

certain symptoms. While there are many possible valid trial designs, if we want to learn about effects on preterm birth throughout pregnancy, we might imagine recruiting a group of participants at varying stages of pregnancy, and randomize them to develop COVID-19 or not. (Ideally we would recruit only people who don't know each other, so we don't risk that someone with COVID-19 infects someone who was randomized not to have it.) Then we would compare: Of the group that got COVID-19 at 22 weeks of pregnancy, how many delivered preterm, compared to the group that enrolled in the trial at 22 weeks but were randomized to stay COVID-free? How about those who enrolled and were randomized at 32 weeks? The COVID-negative group in the latter comparison would be expected to have a lower total risk of preterm delivery compared to the COVID-negative group enrolling at 22 weeks, because they have 10 fewer weeks in which to deliver, but we may find that the COVID-19 group at 32 weeks has a higher risk of preterm birth than both their comparison COVID-negative group at 32 weeks *and* the COVID-positive group at 22 weeks, if COVID-19 increases the risk of preterm birth and has a stronger effect the later in pregnancy it occurs.

In this first chapter I implicitly use the principles of target-trial emulation to specifically assess not only week of infection but also COVID-19 severity and spontaneous vs. induced preterm delivery. In addition, I describe two other complementary epidemiologic analyses that target the same question.

In **Chapter 2** I consider how to design and analyze a trial comparing treatment strategies for recurrent prostate cancer. Previous studies, both experimental and observational, have compared two strategies in this scenario: treat with hormone therapy immediately, or delay for several years. The choice has implications for quality of life as well as possibly survival. However, there is not yet sufficient evidence for recommending one strategy over the other. Unlike the previous chapter, in which COVID-19 was "assigned" at one point in time, these are *sustained* treatment strategies in which treatment is given or withheld over a period of time.

In this chapter I consider treatment strategies for which the exact time at which therapy will be initiated is not known at baseline. Instead, time of initiation depends on disease progression, as measured by the time in which it takes a biomarker – prostate-specific antigen, or PSA – to double. A shorter doubling time indicates that the cancer is growing more quickly, and therefore treatment may be more urgent. By assigning patients with recurrent prostate cancer to different doubling time thresholds at which to start treatment, we can assess whether earlier treatment – in the sense of disease progression, and not necessarily time – can improve survival. Such a treatment strategy would preserve quality of life for as long as possible and perhaps avoid treatment altogether for those whose cancer is slow-growing.

A treatment strategy that involves giving or withholding treatment over a period of time, instead of just once at baseline, requires special care to be sufficiently well-defined. For example, if our treatment strategy requires initiating treatment when a certain threshold of PSA doubling time is reached, but if PSA is never measured, no one would ever get treatment no matter their assigned threshold. Additional components such as this – monitoring of symptoms and biomarkers, a grace period for transitioning to treatment, follow-up to ascertain outcomes – must be fully specified in order to design a trial for a sustained treatment strategy. In addition, when participants don't follow their assigned treatment strategy (which is always the case when emulating a trial in observational data, where no real-life assignment occurred), to compare treatment strategies we must measure and properly account for time-varying confounders, or characteristics that affect both adherence to treatment and survival.

In this chapter, I describe the components necessary to fully specify a target trial for a sustained treatment strategy and specify them for a target trial to compare PSA doubling time thresholds. I also describe two ways to analyze the data: the parametric g-formula and inverse-probability weighted dynamic marginal structural models. Finally, I implement both methods using data from an observational study of prostate cancer patients from clinics around the U.S.

Even the best attempts to use observational data to answer causal questions about human health can fall short. In **Chapter 3**, I consider the bias jointly inflicted on observational studies by unmeasured confounding, selection, and misclassification.

Confounding describes the fact that characteristics that make it more likely for someone to be exposed to or treated with something also might make it more likely for them to experience an outcome, a natural consequence of the fact that most exposures are not randomized. By measuring and properly adjusting for confounders we can mimic randomization – as when we emulate a target trial. When confounders are unmeasured (because, for example, it might be difficult to conceptualize, much less measure, all the structural, social, and behavioral characteristics that lead someone to have a particular diet or take a particular medication), we might want to know whether we would still see an association – the causal effect – between the exposure and outcome of interest, had we been able to take the unmeasured confounding into account.

Similarly, selection bias can result when there are unmeasured characteristics associated both with selection into a study and with the exposure and outcome of interest. Differences in the people participating and not participating in a study – or those who are and are not included in the analytic sample – may be due to dropout, missing data, time and energy available for volunteering, or personal motivations to participate that align with the study's goals. When the participants and non-participants also have different probabilities of exposure and levels of underlying risk of the outcome, estimates of the causal exposure-outcome relationship will be biased.

Finally, measurement error occurs when people are misclassified as having some characteristic when they truly don't, or vice versa. Particularly harmful is differential misclassification, e.g., when people with the exposure are more likely to be misclassified as having the outcome (when they really didn't) than those who were unexposed. The same can occur when misclassification of the exposure is associated with the outcome. For example, when reporting exposures that occurred during pregnancy, someone who delivered prematurely may be more likely to falsely assume they had COVID-19 (based on recall of symptoms, in the absence of a test) than someone

with the same symptoms who delivered a healthy baby at term. This can make it appear that COVID-19 is associated with an increased risk of preterm delivery.

When one or more of these biases threaten the validity of an epidemiologic study, it is worthwhile to conduct additional analyses to assess the robustness of the study's findings. In other words, if we find in the data that an exposure and outcome are apparently associated, would that result still stand if we had been able to account for the possibly biasing factors? One method for conducting such "sensitivity analyses" is assessing the maximum amount of bias that could have been produced under certain conditions. In this chapter, I derive the bound for the bias of an observed risk ratio subject to unmeasured confounding and selection bias and differential misclassification. This bound can be used to compute the minimum value of the causal risk ratio under bias parameters that define the strength of each bias. In other words, we can answer the question: "At worst, how far off is the observed risk ratio from the causal risk ratio, if these biases were of this strength?" I show how to apply this bound to two published studies concerning exposures in pregnancy and offspring health.

In conclusion, this dissertation presents and demonstrates tools and methods to strengthen the validity of using observational data to answer important questions about population health and infer causality with greater confidence.

1

Timing and severity of COVID-19 during pregnancy and the risk of preterm birth

Severe coronavirus disease 2019 (COVID-19) has been associated with preterm delivery. However, previous estimates of risk of preterm delivery after COVID-19 are often subject to selection bias and do not distinguish between infection early vs. late in pregnancy, nor between spontaneous vs. induced preterm delivery. Pregnant and recently pregnant people who were tested for or clinically diagnosed with COVID-19 during pregnancy enrolled in an international internet-based cohort study between June 2020 and March 2021. Using several analytic approaches to minimize biases, we compared the risk of preterm delivery (overall, spontaneous, and induced) among those with and without COVID-19. We also considered different levels of disease severity and timing of infection. There were 14,167 participants from 67 countries eligible for our study, of whom 5,857 had completed their pregnancies and reported delivery information; the remainder were censored at the time of their last follow-up. Participants with COVID-19 before 20 weeks' gestation (n = 2,630) had no increased risk of preterm delivery compared to those testing negative (n = 8,557), with adjusted risks of 9.9% (95% CI 8.1, 12.0) vs. 9.8% (9.2, 10.5). Mild or moderate COVID-19 later in pregnancy was also not associated with preterm delivery. In contrast, severe COVID-19 after 20 weeks' gestation (n = 215) led to an increase in preterm delivery compared to mild or moderate disease (n = 2,350). For example, the estimated risk ratio for severe COVID-19 at 35 weeks was 2.9

This chapter was co-authored with Camille Y. Dollinger, Tyler J. VanderWeele, Diego F. Wyszynski, and Sonia Hernández-Díaz.

(2.0, 4.1); the corresponding risk ratios for induced and spontaneous preterm delivery were 3.6 (2.0, 6.8) and 2.4 (1.3, 3.9), respectively. This elevated risk was primarily due to an increase in induced preterm deliveries, including Cesarean sections due to maternal illness, although an increase in spontaneous preterm delivery was also observed. This study improves upon previous research by providing gestational-age-specific estimates of risk and relative risks that use appropriate comparison groups.

1.1 Introduction

Coronavirus disease 2019 (COVID-19) has proven uniquely harmful to certain populations, including the elderly and individuals with various comorbidities.[1,2] However, its effects on the pregnant population have been less easily discerned. Early studies suggested an elevated risk of preterm birth among pregnant people with COVID-19 at delivery,[3–7] but were limited by small samples from single hospitals, little variability of disease severity or timing of infection during pregnancy, failure to account for pregnancies that continue beyond the study period, and lack of valid comparison groups.

When gestational age at infection is not considered, "immortal time bias" may reduce, negate, or reverse any effect on prematurity.[8] In addition, although the daily rate of preterm delivery increases as week 37 approaches, the total risk of preterm delivery declines over the course of pregnancy due to the shrinking window of time in which to deliver before term, making estimates of risk after infection at different gestational ages difficult to interpret. Most studies have ignored the longitudinal nature of pregnancy and examined the risk of preterm birth associated with COVID-19 at delivery, neglecting to consider gestational age at infection. Finally, while associations between severe COVID-19 and preterm delivery may reflect biological effects of the viral infection or the immunological response, they may also result from medically induced delivery based on health concerns.[5] No study has considered the timing of infection during pregnancy, the severity of disease, the indication of prematurity, and the methodological issues simultaneously.

Using data from a large, international pregnancy cohort, we investigated whether COVID-19 increased the risk of preterm birth. We used multiple analytic approaches to disentangle the role of severe disease and the timing of infection during pregnancy, allowing us to estimate gestational-age-specific risks of preterm delivery after mild, moderate, and severe COVID-19 throughout pregnancy.

1.2 Methods

Cohort

The International Registry of Coronavirus Exposure in Pregnancy (IRCEP) began enrollment in June 2020 for English-speaking participants and has since opened in 9 additional languages; enrollment is ongoing. Enrollees must be pregnant or within 6 months of end of pregnancy and must have had a test for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection or a clinical diagnosis from a healthcare provider of COVID-19 during pregnancy. A valid mobile phone number and internet access are required for enrollment, and participants are asked to submit photos of their test results and delivery records with identifying details removed. Other information is collected via several online survey modules covering demographics, reproductive and health history, COVID-19 symptoms and treatments, and pregnancy and birth outcomes. The Institutional Review Board of the Harvard T.H. Chan School of Public Health approved this study (IRB20-0622).

SARS-CoV-2 infection and COVID-19

At enrollment, participants reported the date, type (nose/throat swab for PCR or blood test for antibodies), and result (positive, negative, inconclusive) of a SARS-CoV-2 test during pregnancy. Participants who were pregnant at enrollment recorded any additional tests during pregnancy on monthly surveys. Clinical diagnoses, symptoms, and treatments of COVID-19 were also reported by all enrollees. We defined COVID-positive participants as those with a positive test or a clinical diagnosis confirmed by a healthcare provider. We considered date of infection to be at symptom onset, if symptoms were reported, or at the time of a positive PCR test in asymptomatic participants. COVID-negative participants were those reporting only negative test(s) at enrollment and no clinical diagnosis of COVID-19.

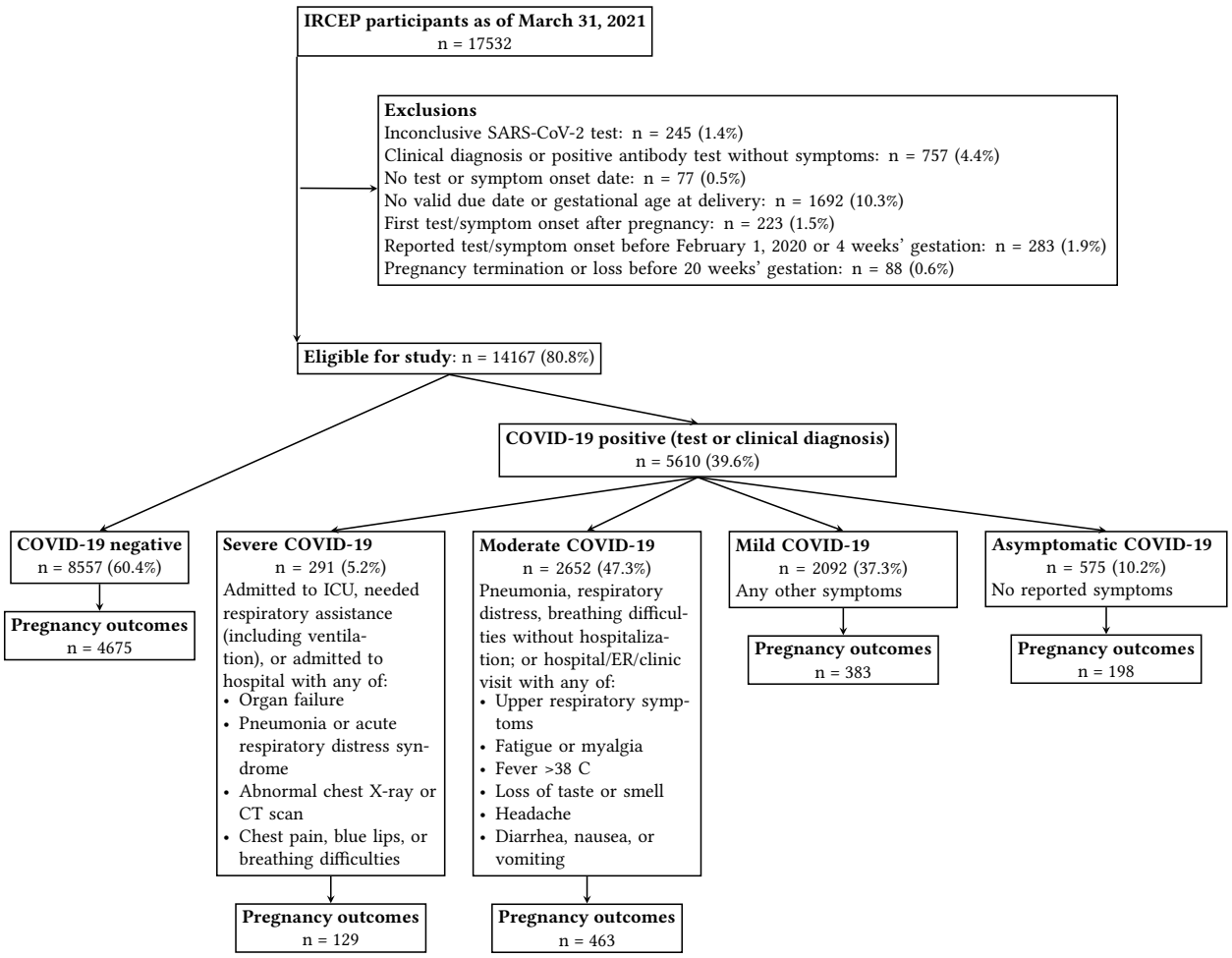

Figure 1.1: *Flowchart of participant eligiblity from the International Registry of Coronavirus in Pregnancy (IRCEP) and classification by COVID-19 severity.*

Following clinical guidelines for classifying COVID-19 severity,[9] we considered anyone who was admitted to the intensive care unit (ICU), needed respiratory assistance (including ventilation or extracorporeal membrane oxygenation (ECMO)), or was hospitalized with reported organ failure, acute respiratory distress syndrome, pneumonia, an abnormal chest X-ray or CT scan, or indications of significant lung involvement to have had severe disease (Figure 1.1). Moderate infections were those with lesser lung involvement or other symptoms that resulted in use of health care outside of the home. Participants with other symptoms were considered mild, and those without reported symptoms, asymptomatic.

Gestational age and delivery outcomes

Participants reported due dates as determined by last menstrual period and by ultrasound when available. Those who were pregnant at enrollment additionally reported date of last menstrual period. Date of delivery, spontaneous pregnancy loss, or termination was reported upon pregnancy completion (at baseline for those who joined postpartum), as was gestational age at end of pregnancy. We used these data to determine gestational age at COVID-19 symptom onset and test date(s), enrollment, and end of pregnancy, in some cases re-contacting participants for clarification when dates were inconsistent.

Preterm birth was defined as delivery before 37 weeks' gestation. Participants also reported mode of delivery and reason for any Cesarean section (C-section). We classified C-sections by whether they were elective or indicated by prior health history, indicated by positioning or size of the fetus, associated with COVID-19 (due to maternal illness or precautionary), due to lack of labor progress, or due to fetal or maternal/placental complications. We also collected information on preterm labor and premature rupture of membranes. We considered preterm delivery to be spontaneous if either spontaneous preterm labor or premature rupture of membranes was reported, and medically induced otherwise.

Study sample

IRCEP participants as of March 31, 2021 who had not incurred early pregnancy loss (<20 weeks) or termination were eligible for our study. We excluded those who reported COVID-19 before February 1, 2020 due to concerns about reporting error, as well as those for whom we could not estimate gestational age. In addition, we excluded those who reported no symptoms and a positive blood (antibody) test, since we did not know when the infection occurred, or who had inconclusive test results. In a sensitivity analysis we additionally excluded those who had received a clinical diagnosis with no positive test.

Participants who had not provided delivery information were excluded from analyses involving mode of delivery. In addition, we excluded from some analyses those who were missing data on baseline covariates. As a sensitivity analysis, we used multiple imputation to impute missing baseline covariates.

Statistical analysis

We compared baseline characteristics and unadjusted risk of preterm delivery and delivery type by COVID-19 status and severity and between those with and without outcome information. Because asymptomatic participants' positive test dates were closely linked to their delivery dates (likely due to routine screening at delivery), artificially increasing the apparent risk of preterm delivery among asymptomatic infections close to 37 weeks, we excluded them from the remaining analyses.[10]

Multivariable regression

Among the participants whose pregnancies had ended in live or still birth, we regressed an indicator of preterm birth on an indicator of symptomatic COVID-19 to estimate the relative risk of preterm birth after COVID-19 at any time vs. never before 37 weeks of pregnancy. Participants with COVID-19 onset after 37 weeks' gestation were excluded, as they were no longer at risk for preterm birth. We also excluded participants with last menstrual periods within 45 weeks prior to the analysis date (a cutoff we varied in sensitivity analyses), to allow sufficient time for term deliveries and avoid over-inclusion of shorter pregnancies. We fit the model using log-linear regression (log-Poisson regression with robust standard errors due to convergence problems with log-binomial regression), adjusting for possible baseline confounding by continent (Africa, Asia, Europe, North America, South America), maternal age (years), pre-pregnancy BMI (kg/m^2), parity (primi-/multiparous), race/ethnicity (Asian, Black, Latina, White, mixed, other), pre-existing condition (chronic diabetes, asthma, cardiovascular disease, or autoimmune disease), healthcare

coverage (yes/no), and reason for testing (symptoms, contact tracing, surveillance, other/not tested). To assess risks specifically due to severe COVID-19, we then restricted the sample to COVID-positive participants and estimated the relative risks of severe and moderate compared to mild disease. Finally, we fit multinomial logistic regression models for a three-leveled outcome (spontaneous preterm, induced preterm, and term delivery) in order to estimate separate odds ratios for each of spontaneous and induced preterm, relative to term delivery.

Accounting for gestational age: Time-to-delivery model

Because longer pregnancies allow for more opportunity for infection, even if COVID-19 did not affect risk of preterm delivery, longer pregnancies could appear more likely exposed to the disease. We therefore matched exposed and reference groups on gestational age-specific start of follow-up (time zero). In addition, the risk of preterm birth differs by week of gestation even in the absence of infection. We therefore considered gestational age-specific risks. Specifically, we asked the question, "What is the risk of preterm delivery in pregnancies affected by COVID-19 at week x of gestation, and how does it compare to the risk in pregnancies that are uninfected but ongoing at week x?"

For every week of gestation through week 36 (the last week in which a pregnancy is at risk of preterm delivery), we selected the individuals whose infection occurred that week and a comparison group made up of all of the COVID-negative participants whose pregnancies were still ongoing at that time (including before enrollment in IRCEP). COVID-negative participants could appear in repeated comparison groups until they delivered or were censored. Participants who had not yet delivered at the time of analysis, or whose delivery date was unknown (i.e., lost to follow-up), were censored at the last known gestational week at which we knew their pregnancy was ongoing. Together the symptomatic positive and the test negative individuals made up week-specific (time zero-specific) subcohorts, within which we computed the daily probability of delivery from the gestational week at time zero through the rest of gestation. We estimated

these probabilities separately for the negative, mild, moderate, and severe groups, using only the observations that were still non-censored by that day. Assuming that censoring is independent of delivery week within each COVID-19 group, the probability of preterm delivery is 1 minus the cumulative product of the probability of not delivering each day up until week 37, analogous to a Kaplan-Meier estimator.

To account for confounding and non-random censoring, we estimated the probability of delivering on a given day with a pooled logistic regression model conditional on the covariates previously described, as well as a flexible function of gestational age (cubic splines) and terms for COVID-19 group (negative, mild, moderate, severe). Because there were relatively few COVID-positive individuals in any time zero-specific subcohort, we increased precision by fitting a model pooled over the participants in all subcohorts, adding cubic splines for time zero, a term for time since infection, and product ("interaction") terms for COVID-19 group and time since infection.

To compute risks of preterm delivery, we predicted probabilities from the model under each condition at each time zero—mild, moderate, or severe infection, or no infection but ongoing pregnancy—for each of the COVID-negative individuals remaining that week, then averaged over the estimated individual risks of delivery before week 37. We combined estimates for early pregnancy (through week 20) as there were no deliveries in those weeks and the number of infections in each was small. Total risks for any infection (i.e., COVID-19 positive) were computed by combining the risks for mild, moderate, and severe disease, weighted by the overall proportion of COVID-positive participants at each level of severity.

We then partitioned our estimates of absolute risk of preterm delivery from the time-to-event model into spontaneous vs. induced preterm. To do so, we fit a logisitic regression model for spontaneous delivery among all preterm deliveries, conditional on gestational age at delivery, COVID-19 severity, weeks since infection, continent, pre-pregnancy BMI, parity, and race. We estimated risk of spontaneous preterm as the probability predicted from that model multiplied by risk of any delivery.

We computed 95% confidence intervals for the risk estimates, and for risk differences and ratios computed from these risks, using the non-parametric bootstrap with 1000 replicates (which accounted for the fact that COVID-negative individuals could contribute to multiple subcohorts).

Robustness to unmeasured confounding: Case-time-control design

The previous analyses assume that the measured confounders were sufficient to control confounding; to reduce risk of bias by unmeasured between-person time-fixed confounders, we additionally conducted a within-person analysis using a case-time-control design.[11,12] Although both the study population and the parameter being estimated are different from each of the other analyses, an association in this design would support the presence of effects. If there is an acute, transient effect of COVID-19 on preterm birth, *among* people who delivered preterm (i.e., those susceptible to prematurity), COVID-19 will more likely occur during the period in which it affected delivery timing (i.e., presumptively the weeks prior to delivery) than any other period in pregnancy. We therefore compared, among preterm births (cases), the probability of COVID-19 in the 30 days preceding delivery to the probability of COVID-19 in a reference period 120-90 days prior to delivery, when we hypothesized it is less likely to affect delivery. However, infection is not equally likely in every week of pregnancy, even if it does not affect week of delivery, so we made the same comparison among term births (controls) to estimate the time trend. We matched each preterm case with one or more term controls on calendar year and month of due date. Among the controls we compared the odds of COVID-19 in the gestational age windows that corresponded to those of their matched cases, then divided out this time effect from the total effect among the cases. This process is equivalent to fitting a conditional logistic regression for exposure with indicators for case/control status, risk/reference period, and their interaction, among a dataset with a row for each observation in the risk and reference period. As a sensitivity analysis, we repeated the analysis using a range of windows for the risk and reference periods. To assess differences by

COVID-19 severity, we repeated the analysis using an indicator of severe disease as the exposure, and separately with an indicator of mild/moderate disease.

1.3 Results

Sample

Of 14,167 eligible IRCEP participants from 67 countries, 60.1% joined the registry while pregnant (Figure 1.1). Compared to individuals testing negative, those with COVID-19 were of similar ages (30.1 vs. 30.5, among positive and negative, respectively) and had similar pre-pregnancy BMI (26.7 vs. 27.2). However, participants with positive tests were more likely to be from South America (37.0% of positive vs. 15.7% of negative) and had fewer preexisting conditions (11.9% vs. 15.2%), the latter likely reflecting more screening of high-risk individuals (Table 1.1).

Table 1.1: *Descriptive characteristics (n (%)) of eligible International Registry of Coronavirus Exposure in Pregnancy (IRCEP) participants enrolled June 2020–March 2021.*

	COVID-19 negative N = 8,557	COVID-19 positive N = 5,610	Total N = 14,167
Enrollment			
Prospective (during pregnancy)	3,973 (46%)	4,536 (81%)	8,509 (60%)
Retrospective (after pregnancy)	4,584 (54%)	1,074 (19%)	5,658 (40%)
Prospective enrollees			
Follow-up data	237 (6.0%)	155 (3.4%)	392 (4.6%)
Gestational age at enrollment[a]	27 (18, 35)	25 (17, 33)	26 (17, 34)
Gestational age at symptom onset/text[a]	21 (12, 30)	19 (10, 27)	20 (11, 28)
Retrospective enrollees			
Follow-up data	4,438 (97%)	1,018 (95%)	5,456 (96%)
Gestational age at enrollment[a]	49 (44, 54)	47 (43, 53)	49 (44, 54)
Gestational age at symptom onset/test[a]	38.4 (36.9, 39.4)	33.6 (27.6, 37.6)	38.0 (35.4, 39.3)

Table 1.1: *Descriptive characteristics (n (%)) of eligible International Registry of Coronavirus Exposure in Pregnancy (IRCEP) participants enrolled June 2020–March 2021. (continued)*

	COVID-19 negative N = 8,557	COVID-19 positive N = 5,610	Total N = 14,167
COVID-19 severity			
Negative	8,557 (100%)	–	8,557 (60%)
Asymptomatic	–	575 (10%)	575 (4.1%)
Mild	–	2,092 (37%)	2,092 (15%)
Moderate	–	2,652 (47%)	2,652 (19%)
Severe	–	291 (5.2%)	291 (2.1%)
COVID-19 diagnosis/test type			
Negative	8,557 (100%)	–	8,557 (60%)
Positive by antibodies only	–	531 (9.5%)	531 (3.7%)
Positive by throat/nose swab	–	4,465 (80%)	4,465 (32%)
Positive clinically only	–	614 (11%)	614 (4.3%)
Reason for COVID-19 test			
Symptoms	1,222 (14%)	4,060 (72%)	5,282 (37%)
Contact tracing/risk zone travel	1,672 (20%)	970 (17%)	2,642 (19%)
Surveillance (healthy)	2,439 (29%)	224 (4.0%)	2,663 (19%)
Other/none	3,223 (38%)	355 (6.3%)	3,578 (25%)
Age[a]	31.0 (27.0, 34.0)	30.0 (27.0, 34.0)	31.0 (27.0, 34.0)
Healthcare coverage	6,629 (89%)	3,685 (85%)	10,314 (87%)
Pre-existing condition	1,101 (15%)	478 (12%)	1,579 (14%)
Primiparous	3,337 (46%)	1,691 (42%)	5,028 (44%)
Pre-pregnancy BMI			
<25	3,187 (47%)	1,778 (48%)	4,965 (47%)
25-30	1,757 (26%)	1,019 (28%)	2,776 (26%)
≤ 30	1,841 (27%)	894 (24%)	2,735 (26%)
Continent			
Africa	376 (4.4%)	237 (4.2%)	613 (4.3%)
Asia	488 (5.7%)	411 (7.3%)	899 (6.3%)
Europe	3,079 (36%)	1,304 (23%)	4,383 (31%)
North America	3,266 (38%)	1,585 (28%)	4,851 (34%)
South America	1,346 (16%)	2,073 (37%)	3,419 (24%)

[a] Median (interquartile range)

Prospective participants, who had a mean gestational age of 25.3 weeks at enrollment (median 26.1) were much more likely to report a positive test than retrospective participants (53.3% vs. 19.0% positive, respectively), reflecting a pattern of COVID-19 screening near delivery resulting in more negative (or asymptomatic positive) tests (Figure A.1), while a larger proportion of testing during pregnancy was triggered by symptoms or for other non-pregnancy-related reasons. On average, retrospective participants enrolled at 10.5 weeks after end of pregnancy (median 10.0).

At the time of analysis, participants had reported information about 5,820 live births and 37 stillbirths (Table 1.2). Of those who joined while pregnant, 4.6% had already provided outcome data, 36.4% were less than 42 weeks' gestation at the time of this analysis or reported still being pregnant on a monthly survey, and 59.0% were more than 42 weeks but had not yet provided outcomes (i.e., presumed lost to follow-up). Of those who joined after pregnancy, 96.6% had provided at least some outcome data. Participants with positive tests or diagnoses were about as likely to provide outcome data after pregnancy completion (or still be pregnant) as those testing negative, both among prospective pregnancies (41.3% of positive vs. 40.6% of negative) and retrospective pregnancies (95.0% of positive vs. 96.9% of negative). Participants providing outcome data were equally likely as those who did not to have preexisting conditions (13.8%), and age and pre-pregnancy BMI were similar across groups (Table A.1). Those with outcome data were more likely from Europe or North America than other continents, were slightly more likely to be primiparous (46.2% vs. 43.3%) and to have healthcare coverage (88.8% vs. 84.1%). Among the 5,034 pregnant individuals with symptomatic COVID-19 (89.7% of positive participants), we classified 291 as severe, 2,652 moderate, and 2,091 mild (Figure 1.1); 976 had available outcomes, with 30.2%, 11.7%, and 9.1% preterm deliveries in the severe, moderate, and mild groups, respectively. Of preterm births in the severe group, 82.1% were C-sections, compared to 61.8% in the mild and moderate groups combined. Among the severe group, 54.8% of preterm C-sections specifically mentioned illness from or precautions due to COVID-19 as the reason for the procedure, compared to 7.8% in the mild/moderate groups (Figure 1.3). Overall, 48.7% of preterm deliveries in the severe group were emergency (vs. planned) C-sections following COVID-19 complications, fetal distress, acute

Table 1.2: *Delivery outcomes among IRCEP participants.*

	Negative N = 4,675	Asymptomatic N = 198	Mild N = 383	Moderate N = 463	Severe N = 129
Preterm delivery	411 (8.8%)	22 (11%)	35 (9.1%)	54 (12%)	39 (30%)
Type of preterm delivery					
Indicated	131 (2.8%)	10 (5.1%)	12 (3.1%)	19 (4.1%)	16 (12%)
Spontaneous	280 (6.0%)	12 (6.1%)	23 (6.0%)	35 (7.6%)	23 (18%)
Premature rupture of membranes	136 (2.9%)	6 (3.1%)	11 (2.9%)	17 (3.7%)	6 (4.8%)
Preterm labor	237 (5.1%)	12 (6.1%)	14 (3.7%)	39 (8.5%)	14 (11%)
Induced labor	1,891 (42%)	67 (36%)	123 (34%)	172 (39%)	44 (38%)
Cesarean-section	1,880 (41%)	95 (50%)	168 (45%)	218 (49%)	78 (64%)
Primary reason for C-section					
Elective or health history	750 (16%)	33 (17%)	69 (19%)	87 (20%)	26 (21%)
Labor not progressing	459 (10%)	19 (10%)	33 (8.9%)	45 (10%)	6 (5.0%)
Positioning or size	321 (7.0%)	17 (9.0%)	29 (7.8%)	26 (5.8%)	2 (1.7%)
Maternal/placental problems	180 (3.9%)	12 (6.3%)	20 (5.4%)	24 (5.4%)	8 (6.6%)
Fetal distress, cord or other problems	122 (2.7%)	1 (0.5%)	7 (1.9%)	9 (2.0%)	5 (4.1%)
COVID precautions	20 (0.4%)	12 (6.3%)	5 (1.3%)	12 (2.7%)	11 (9.1%)
COVID complications	0 (0%)	0 (0%)	1 (0.3%)	5 (1.1%)	18 (15%)
Unknown	28 (0.6%)	1 (0.5%)	4 (1.1%)	10 (2.2%)	2 (1.7%)

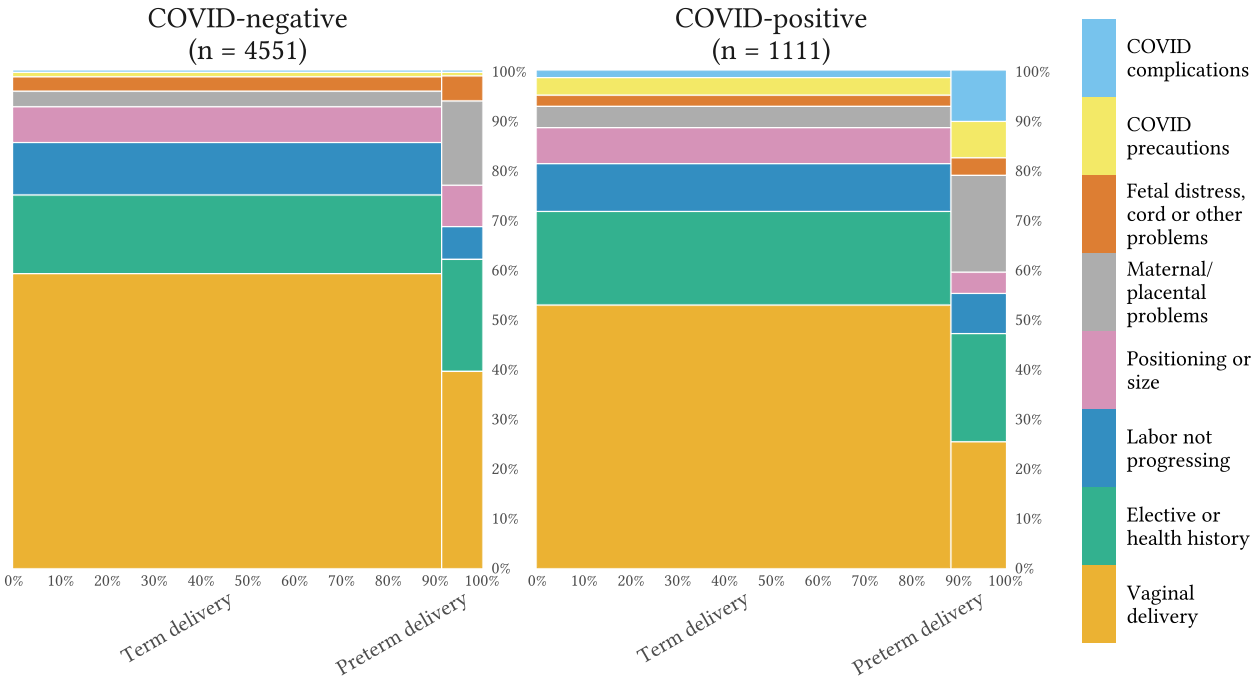

Figure 1.2: *Mode of delivery among IRCEP participants, stratified by COVID-19.*

maternal or placental problems, or labor not progressing, compared to 30.3% of mild and moderate preterm deliveries.

Multivariable regression

Completed pregnancies with COVID-19 exposure any time before 37 weeks were 1.3 (95% CI 1.0, 1.7) times as likely to deliver preterm as those testing negative, and those with severe disease 2.5 (1.6, 3.9) times as likely as with mild disease (Table 1.3). The risk ratio comparing moderate to mild disease was 1.1 (0.7, 1.7). From the multinomial logistic regression, COVID-19 during pregnancy was associated with 1.1 (0.7, 1.6) times the odds of spontaneous preterm, and 2.1 (1.2, 3.7) times the odds of induced preterm, relative to term birth. Odds ratios for severe vs. mild disease were 2.6 (1.3, 5.3) and 6.0 (2.4, 14.9) for spontaneous and induced preterm delivery, respectively (Table 1.3).

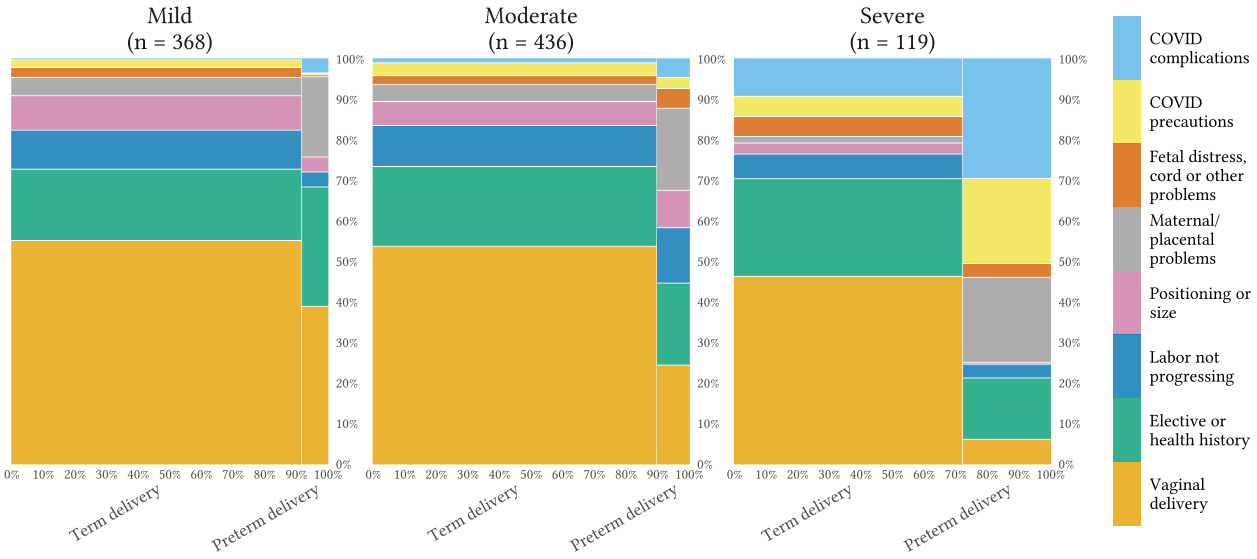

Figure 1.3: *Mode of delivery among symptomatic IRCEP participants with COVID-19, stratified by severity.*

Time-to-delivery

Unadjusted gestational-age-specific absolute risks of preterm delivery varied depending on week of infection. Differences appeared to emerge after infection in the third trimester (Figure A.2). Adjusted absolute risks of preterm delivery also varied by gestational age (Figure 1.4); for example, the risk of preterm delivery after COVID-19 before 20 weeks of pregnancy was 9.9% (8.1, 12.0) and 9.8% (9.2, 10.5) among pregnancies that were ongoing but not infected at that time, compared to 7.5% (6.5, 8.5) and 6.9% (6.4, 7.3), respectively, at 35 weeks. Risk was 9.1% (6.1, 13.7) after severe disease before 20 weeks and 19.4% (13.7, 27.6) after severe disease at 35 weeks. We combined mild and moderate disease for the remaining analyses, as risks were essentially identical (estimated separately in a sensitivity analysis, Figure A.5). Risk ratios comparing severe to mild/moderate disease at 20 and 35 weeks were 0.9 (0.7, 1.3), and 2.9 (2.0, 4.1), respectively. Table 1.3 contains absolute risks and risk ratios for additional weeks, and Table A.4 risk differences. Compared to mild/moderate disease, risks were higher for both spontaneous and induced preterm delivery after severe COVID-19 (Figure 1.5). For example, after infection at 35 weeks, risk ratios were 2.4 (1.3, 3.9) and 3.6 (2.0, 6.8) for spontaneous and induced preterm delivery, respectively (Tables A.2 and A.3).

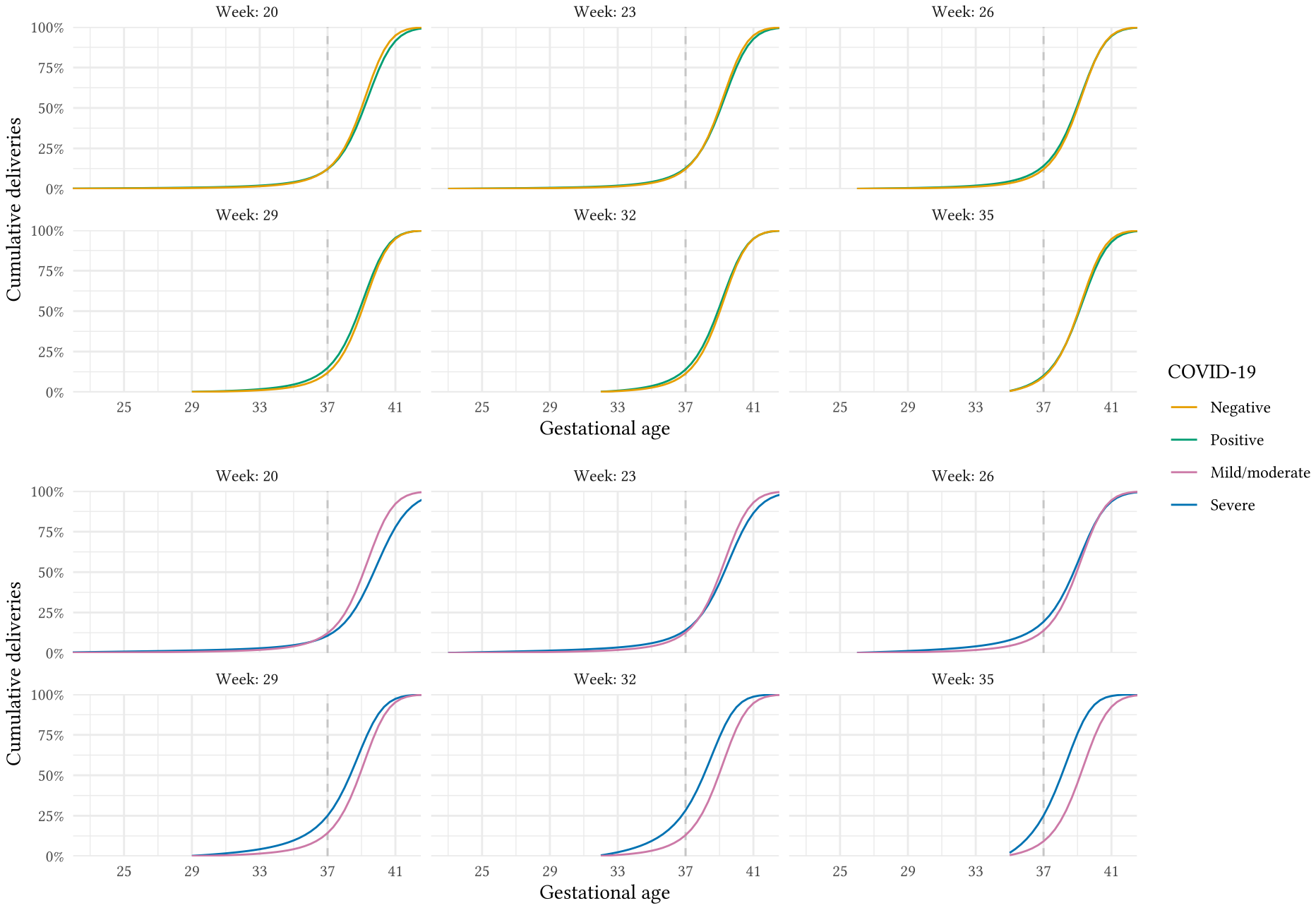

Figure 1.4: *Standardized cumulative probabilities of delivery after COVID-19 in selected weeks of gestation. COVID-19 negative individuals in a given week are those who are still pregnant at that week. Week 20 refers to all infections at or before week 20.*

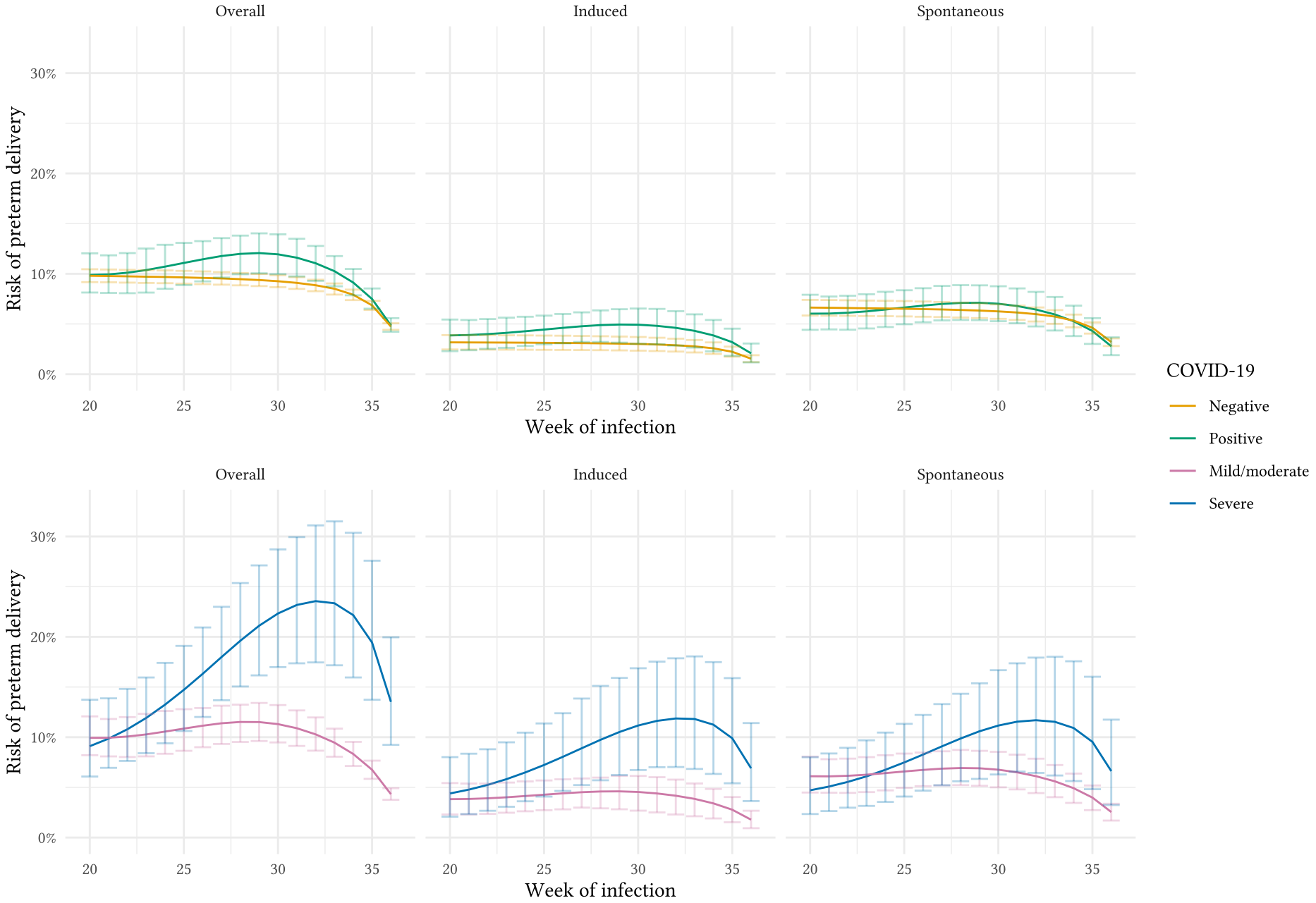

Figure 1.5: *Risks of preterm delivery from infection to end of pregnancy, according to week of infection and COVID-19 severity.*

Case-time-control

Cases had higher odds of having had COVID-19 in the month prior to their preterm deliveries compared to 3-4 months prior (odds ratio of 1.5 (0.8, 2.9)), accounting for time trends in exposure (Table 1.3). In addition, they had 3.9 (0.7, 21.2) times increased odds of having had severe COVID-19 in the month prior to their preterm deliveries compared to 3-4 months prior, but only 0.9 (0.4, 1.9) times the odds of a mild or moderate infection (Table 1.3).

Sensitivity analyses

Estimates from the log-linear analyses were almost identical when we imputed baseline covariates, as well as when we excluded participants who had only a clinical diagnosis of COVID-19 and no positive test (Table A.5). Varying the cutoff for time since last menstrual period in the log-linear analyses also resulted in almost identical estimates (Table A.6), as did excluding participants whose negative tests were in the final two weeks of delivery (Figure A.4). Estimating risks separately for mild and moderate disease had almost identical results as combining the groups (Figure A.5). Restricting our sample to only North American participants resulted in slightly larger risk ratios for both severe vs. mild and moderate vs. mild disease (Table A.6). When we restricted the analysis to prospective participants only, there was a smaller difference in risk between severe and mild/moderate disease (Figure A.5), likely because of the relatively small number of individuals joining after severe disease but before delivery late in pregnancy. Varying timing and duration of the risk period in the case-time control analysis confirmed our hypothesis that risk due to infection occurred in the month prior to delivery, with particularly strong risk in the 2 weeks preceding delivery (Figure A.3).

Table 1.3: *Estimates of risk of preterm birth from the various models and comparisons across levels of COVID-19. Gestational age-specific results are presented by week.*

Model	Standardized risks by COVID-19[a]				Risk ratios[a]	
	Negative	Positive	Mild/moderate	Severe	Positive vs. negative	Severe vs. mild or mild/moderate
Log-linear regression					1.3 (1.0, 1.7)	2.5 (1.6, 3.9)
Multinomial regression (Induced)[b]					2.1 (1.2, 3.7)	6.0 (2.4, 14.9)
Multinomial regression (Spontaneous)[b]					1.1 (0.7, 1.6)	2.6 (1.3, 5.3)
Case-time-control[b]					1.5 (0.8, 2.9)	3.9 (0.7, 21.2)[c]
Gestational age-specific						
Week 20	9.8% (9.2, 10.5)	9.9% (8.1, 12.0)	9.9% (8.2, 12.1)	9.1% (6.1, 13.7)	1.0 (0.8, 1.2)	0.9 (0.7, 1.3)
Week 21	9.8% (9.2, 10.4)	9.9% (8.1, 11.8)	9.9% (8.1, 11.8)	9.9% (6.9, 13.9)	1.0 (0.8, 1.2)	1.0 (0.8, 1.4)
Week 22	9.7% (9.1, 10.4)	10.1% (8.1, 12.1)	10.1% (8.0, 12.0)	10.8% (7.6, 14.8)	1.0 (0.8, 1.3)	1.1 (0.8, 1.4)
Week 23	9.7% (9.1, 10.4)	10.4% (8.1, 12.5)	10.3% (8.1, 12.3)	11.9% (8.4, 15.9)	1.1 (0.8, 1.3)	1.2 (0.9, 1.5)
Week 24	9.7% (9.1, 10.3)	10.7% (8.5, 12.9)	10.6% (8.3, 12.6)	13.2% (9.4, 17.4)	1.1 (0.9, 1.3)	1.3 (1.0, 1.6)
Week 25	9.6% (9.0, 10.3)	11.1% (8.9, 13.1)	10.9% (8.6, 12.8)	14.7% (10.6, 19.1)	1.1 (0.9, 1.4)	1.4 (1.1, 1.7)
Week 26	9.6% (9.0, 10.2)	11.4% (9.2, 13.3)	11.1% (9.0, 12.9)	16.3% (12.0, 20.9)	1.2 (1.0, 1.4)	1.5 (1.2, 1.9)
Week 27	9.5% (8.9, 10.2)	11.8% (9.6, 13.6)	11.4% (9.3, 13.1)	18.0% (13.7, 23.0)	1.2 (1.0, 1.4)	1.6 (1.3, 2.0)
Week 28	9.5% (8.8, 10.1)	12.0% (10.0, 13.8)	11.5% (9.5, 13.2)	19.6% (15.1, 25.4)	1.3 (1.1, 1.5)	1.7 (1.4, 2.2)
Week 29	9.4% (8.8, 10.0)	12.1% (10.0, 14.0)	11.5% (9.6, 13.4)	21.1% (16.2, 27.1)	1.3 (1.1, 1.5)	1.8 (1.5, 2.4)
Week 30	9.3% (8.7, 9.9)	11.9% (10.0, 13.9)	11.3% (9.5, 13.2)	22.3% (17.0, 28.7)	1.3 (1.1, 1.5)	2.0 (1.5, 2.6)
Week 31	9.1% (8.5, 9.7)	11.6% (9.7, 13.5)	10.9% (9.1, 12.7)	23.2% (17.4, 29.9)	1.3 (1.1, 1.5)	2.1 (1.6, 2.8)
Week 32	8.9% (8.3, 9.4)	11.1% (9.3, 12.8)	10.3% (8.7, 12.0)	23.6% (17.4, 31.1)	1.2 (1.1, 1.5)	2.3 (1.7, 3.1)

Table 1.3: *Estimates of risk of preterm birth from the various models and comparisons across levels of COVID-19. Gestational age-specific results are presented by week. (continued)*

Model	Negative	Positive	Mild/moderate	Severe	Positive vs. negative	Severe vs. mild or mild/moderate
Week 33	8.5% (7.9, 9.0)	10.3% (8.8, 11.8)	9.5% (8.1, 10.9)	23.3% (17.2, 31.5)	1.2 (1.0, 1.4)	2.5 (1.8, 3.4)
Week 34	7.9% (7.4, 8.4)	9.1% (7.9, 10.5)	8.3% (7.1, 9.5)	22.2% (16.0, 30.4)	1.2 (1.0, 1.3)	2.7 (1.9, 3.7)
Week 35	6.9% (6.4, 7.3)	7.5% (6.5, 8.5)	6.8% (5.8, 7.7)	19.4% (13.7, 27.6)	1.1 (1.0, 1.3)	2.9 (2.0, 4.1)
Week 36	4.7% (4.4, 5.1)	4.9% (4.2, 5.6)	4.3% (3.8, 4.9)	13.5% (9.2, 20.0)	1.0 (0.9, 1.2)	3.1 (2.1, 4.7)

[a] Adjusted for continent (Africa, Asia, Europe, North America, South America), maternal age (years), pre-pregnancy BMI (kg/m^2), parity (primi-/multiparous), race/ethnicity (Asian, Black, Latina, White, mixed, other), pre-existing condition (chronic diabetes, asthma, cardiovascular disease, or autoimmune disease), healthcare coverage (yes/no), and reason for testing (symptoms, contact tracing, surveillance, other/not tested).

[b] Odds ratios

[c] Severe vs. other

1.4 Discussion

In a large, diverse pregnancy cohort, we found that severe COVID-19 late in pregnancy may double or triple the probability of preterm delivery, depending on the week of infection, but that increased risk due to milder disease, or earlier in pregnancy, is likely minimal. Much of the effect of severe COVID-19 appears to be due to emergency Cesarean sections and other induced preterm deliveries: compared to 25% of preterm births among COVID-19 test negative individuals, almost 50% of the preterm births associated with severe COVID-19 were delivered by emergency C-section. Of those, half specifically reported COVID-19 as the reason for the procedure. Indeed, some participants in our study who underwent emergency C-section due to severe illness described harrowing delivery experiences, writing, for example: "I was struggling to breathe and could not push"; "I thought I was gonna die"; "I had an emergency Cesarean section while I was in a coma induced by COVID-19, I was 24 weeks pregnant." An effect of severe COVID-19 on iatrogenic preterm delivery is clear.

Nonetheless, a higher proportion of COVID-19-affected pregnancies in our study also reported preterm labor or rupture of membranes, making it impossible to rule out effects on spontaneous preterm delivery. Intrauterine bacterial infection is known to be a major cause of spontaneous preterm delivery,[13] mediated through the innate immune response,[14] and concern about the effects of SARS-CoV-2 infection during pregnancy is justified by evidence from other viral infections.[15,16] Research shows that influenza infection, including 2009 pandemic H1N1 influenza, increases risk of poor birth outcomes, including preterm birth.[17–20] Outbreaks of SARS and MERS, both coronaviruses related to SARS-CoV-2, provided evidence that infection was particularly harmful during pregnancy, with reports of preterm birth, intrauterine growth restriction, and mortality.[21–23]

Case reports, case series, and other initial studies early in the pandemic described outcomes among pregnant people with COVID-19; a number of meta-analyses have since estimated risk of preterm birth pooled from these studies.[24–29] Risk estimates for preterm birth after COVID-19 from meta-analyses range from at least 14% to 61%;[30] with even wider variability across the

individual studies, including geographically and by study size.[27] Apparent differences in risk result from selection of hospitalized patients with severe or critical disease only, exclusion of ongoing pregnancies, and inclusion of pregnancies that have already passed 37 weeks' gestation at symptom onset. Unfortunately, the flaws that prevent interpretation of individual studies and of comparisons between them aren't attenuated through meta-analysis. Indeed, a recent review of systematic reviews on COVID-19 in pregnancy found only one[24] out of 52 was at low risk of bias.[30] Our study improves upon previous estimates of risk by providing gestational-age specific risks among pregnant people with both severe and mild/moderate disease using data from ongoing as well as completed pregnancies.

Other studies, including some based on surveillance data, have also provided valuable comparisons with concurrent[31–34] or historical[35–37] COVID-19-negative pregnancies, or across the spectrum of disease severity.[38–43] Results are mixed, with many[35,38–43] but not all[32–34] providing evidence that any vs. no infection, or more vs. less serious disease, are associated with higher preterm risk. However, the extent to which prior studies have adjusted for confounders or avoided other sources of bias differs. For example, the use of historical reference is questionable given that the pandemic affected health care during 2020 beyond the infection itself; and the lack of adjustment for risk factors for the infection and its severity (e.g., diabetes, obesity, asthma, race/ethnicity) can confound the relative risk estimates. Moreover, studies that define exposure to COVID-19 as "any time during pregnancy" or "any time during third trimester" will be biased by design, given the shorter opportunity for infection in preterm deliveries, and also obscure gestational-age-specific risks. Ultimately, estimates of preterm risk or comparisons of risk between groups need to target the same estimands to be directly comparable; this is particularly difficult for questions surrounding preterm delivery, when timing of the exposure and of sample selection play an important role.

We used three distinct designs to investigate whether COVID-19 affects risk of preterm birth; although the estimates from these analyses are not themselves directly comparable due to differences in the underlying estimands, each addresses shortcomings of other designs. We addressed

confounding both with multivariable adjustment as well as with within-person comparisons using a case-time-control design. In addition, we accounted for several timing-related issues: longer pregnancies are more likely to have been exposed to SARS-CoV-2, risk of preterm delivery depends on gestational age at time zero (infection or reference), and preterm delivery cannot be assessed among participants whose pregnancies are ongoing and under 37 weeks.

Nevertheless, our study has important limitations. Information on tests and gestational age at delivery was self-reported. However, mothers are likely to remember results of their COVID tests in the weeks afterward, as well as their estimated due date and date of delivery. In addition, we have limited clinical measures compared to studies based on medical records or direct clinical observation, limiting our ability to classify COVID-19 cases by severity. We used objective and standard measures of severity (e.g., ICU, ventilation, ECMO) to maximize specificity, at the possible cost of sensitivity (e.g., some hospitalizations may have been precautionary due to pregnancy). This misclassification would tend to bias toward the null; therefore, we may have underestimated associations with severe COVID-19.

Furthermore, while we do not have outcomes on some participants due to ongoing pregnancy, others have been lost to follow-up. In our week-specific risk analyses we were able to use data until the last known week of continued pregnancy under the assumption that loss to follow-up was independent of preterm birth, conditional on covariates. In addition, although we had participants worldwide, some countries were represented more than others, and within-country sampling was not random. While our study is more widely representative than previous studies, there are undoubtedly people who aren't represented, in particular those without sufficient internet access. We may be missing other populations badly affected by the pandemic and who should be prioritized in further research. Although the intensity of the pandemic may differ geographically, biologic effects of COVID-19 may be less likely to vary across the population. However, if the effect is mediated through precautionary early C-sections to avoid transmission or maternal complications during labor, then populations unable or hesitant to conduct these would not see an increased risk of preterm delivery; instead, risk of other periconceptional morbidities may rise.

In conclusion, this study suggests that with respect to preterm birth, prevention of COVID-19 is especially important in the second half of pregnancy. Protective measures should be taken to avoid SARS-CoV-2 infection and symptoms should be closely monitored to avoid disease progression. Much of the increased risk of prematurity due to severe disease is likely iatrogenic, due to urgent delivery in response to maternal or fetal decline. Vaccines could lower the risk of infection, the first step in the causal path towards severe disease, and improved treatments for COVID-19 could lower risk of progression and thus prematurity by reducing indications for delivery; research into these treatments should not exclude pregnant people.

2

Emulation of a target trial with sustained treatment strategies: An application to prostate cancer

Recurrent prostate cancer is generally incurable, although treatment with androgen deprivation therapy can prolong survival. Even in the absence of treatment, however, disease progression may be slow, and negative impacts on quality of life may outweigh possible benefits of initiating treatment directly after recurrence. Previous studies have not determined whether immediate treatment is preferable to deferred treatment, nor have they considered strategies for initiating treatment based on characteristics of prostate-specific antigen (PSA), which can indicate disease progression. We define the protocol for a target trial comparing treatment strategies based on PSA doubling time, in which androgen deprivation therapy is initiated only after doubling time decreases below a certain threshold. Such a treatment strategy means the timing of treatment initiation (if ever) is not known at baseline, and the target trial protocol must explicitly specify the frequency of PSA monitoring until the threshold is met, as well as the duration of treatment. We describe these and other components of a target trial that need to be specified in order for such a trial to be emulated in observational data. We then use the parametric g-formula and inverse-probability weighted dynamic marginal structural models to emulate our target trial in a cohort of prostate cancer patients from clinics across the United States.

This chapter was co-authored with Xabier García de Albéniz, June M. Chan, and Miguel Hernán.

2.1 Introduction

When randomized trials are not feasible or timely, observational data can be used to emulate the randomized trial that, if conducted, would answer the question of interest – the target trial.[44] Observational emulations can result in effect estimates that match those from true randomized trials,[45–48] but these comparisons typically benefit from the protocol of the target trial being explicitly specified.[49] In particular, the treatment strategies under comparison need to be unambiguously described, which may not be a simple task when the strategies are sustained over time.

As an example, consider the question of when to start treatment in people with previously treated prostate cancer who experience a rise in prostate-specific antigen (PSA) without overt metastasis or symptoms.[50] The specification of the treatment strategies includes not only the criteria for both treatment initiation (e.g., PSA greater than some value) and treatment discontinuation (e.g., side effects or a planned intermittent treatment strategy),[51,52] but also the duration of the allowable period to start treatment after the criteria are reached (the grace period), and the frequency of monitoring for those criteria.

Here we describe the components necessary to specify protocols for target trials involving sustained treatment strategies. As an illustration we specify and emulate in observational data a target trial of dynamic strategies for androgen deprivation therapy for recurrent prostate cancer. We illustrate the use of two methods to adjust for time-varying confounding: inverse probability weighting of dynamic marginal structural models[53,54] and the parametric g-formula.[55,56]

2.2 The target trial

We previously emulated a target trial among people with prostate cancer and PSA-only relapse to compare immediate initiation of treatment vs. deferral of treatment.[57] Immediate initiation was defined as androgen deprivation therapy prescription or orchiectomy within three months of the PSA-based relapse, and deferral as a lack of treatment within two years of this relapse or until

evidence of progression. The estimates from our emulation were compatible with those from two subsequent randomized controlled trials,[58,59] which found small differences in all-cause mortality between the two treatment strategies. The 95% confidence intervals for both the observational and randomized effect estimates were very wide.[57,59]

One factor affecting prostate cancer prognosis after biochemical relapse is change in PSA over time.[60] However, the three previous studies[57–59] considered treatment initiation strategies that did not depend on evolving PSA levels. Assigning androgen deprivation therapy only to patients with worsening prognosis might avoid or delay the costs and side effects of treatment[51] for others (who may not benefit). We therefore designed a target trial that assigns treatment only when rapid increases in PSA were observed. The key components of this protocol are summarized in Table 2.1 and below.

Briefly, the trial would include individuals diagnosed with early-stage prostate cancer treated with curative intent who later had evidence of recurrence that was only apparent as a rise in PSA (approximately the same criteria as in the previous studies). Patients would be assigned to one of 37 treatment strategies, each based on a PSADT threshold for treatment initiation (see below). Each eligible individual would be followed from the assignment of the treatment strategy (time zero) until death (the outcome of interest), administrative end of follow-up (10 years after time zero), or loss to follow-up (2 years without contributing clinical data), whichever occurs first. The data from the target trial could be used to estimate both the intention-to-treat effect and the per-protocol effect.[61] The treatment strategies and statistical analysis plan are described in more detail in the following sections.

Target trial: Treatment strategies

The target trial would compare treatment initiation strategies that depend on the individual's rate of change in PSA levels, i.e., the PSA velocity. A common measure of PSA velocity is the PSA doubling time (PSADT), that is, the estimated time over which PSA would double, given

observed values.[62] Specifically, we calculate PSADT from consecutive measurements of PSA at time s and time t as $\frac{\{\log(2\times \text{PSA}_s)-\log(\text{PSA}_s)\}\times\{\text{date}_t-\text{date}_s\}}{\log(\text{PSA}_t)-\log(\text{PSA}_s)} = \log(2)\frac{\text{date}_t-\text{date}_s}{\log\left(\frac{\text{PSA}_t}{\text{PSA}_s}\right)}$ where the difference between measurement dates is in days. The trial would include 37 treatment strategies of the form "Start androgen deprivation therapy the first time PSADT drops below x days," where the threshold x varies from 0 to 360 in increments of 10. A threshold of 0 means that treatment would never be initiated. The treatment duration would be left to be decided by the physician and patient, but this target trial does not allow for intermittent treatment: once treatment is discontinued for longer than one month, it is not to be re-initiated. This description of the treatment strategies is, however, incomplete for the following four reasons:

First, because treatment may be clinically indicated in situations not defined solely by PSADT, the treatment strategies need to specify the situations under which treatment is indicated regardless of PSADT. For example, "Start treatment the first time PSADT drops below x days, *or if a patient shows other signs of progression based on imaging or severe symptoms*."

Second, because immediate initiation of treatment may be unfeasible, we need to also specify the period during which treatment can be started (the grace period). For example, "Start treatment *within the three months following* the first time PSADT drops below x days or the time a patient shows other signs of progression based on imaging or severe symptoms." In practice, randomized trials rarely specify the duration of the grace period because it is understood that treatment will be initiated reasonably soon after randomization. However, specifying the duration of the grace period in the target trial is required in order to emulate it using observational data. Otherwise, it would not be possible to determine whether an individual who initiated treatment, say, 1 year after meeting initiation criteria has data compatible with the protocol of the target trial.

Third, because initiation of treatment during the grace period may follow many patterns (e.g., most people start treatment at the beginning of the grace period, or at the end of the grace period, or uniformly throughout the grace period), we also need to specify the expected rate of treatment initiation during the grace period. For example, "Start treatment *with equal probability* during any

of the three months following the first time PSADT drops below x days, or if a patient shows other signs of progression based on imaging or severe symptoms." Again, actual randomized trials rarely specify the expected rate of treatment initiation during the grace period, but this information is required for the observational emulation, as we discuss below.

Finally, because the initiation of treatment depends on the frequency of measurement of PSA and other characteristics, we also need to specify the intensity of monitoring. For example, "Start treatment with equal probability during any of the three months following the first time PSADT drops below x days, or if a patient shows other signs of progression based on imaging or severe symptoms. *Participants must visit their physician for tests, imaging, and/or symptom assessment in addition to completing surveys at home not less than once every 2 years.*"

Table 2.1: *Description of a target trial to identify the optimal androgen deprivation therapy timing with respect to prostate-specific antigen doubling time, and of its emulation in observational data.*

Components	Target trial	Emulation of trial
Aim	To identify the optimal PSA doubling time at which to begin ADT with respect to 5- and 10-year all-cause mortality after PSA-only relapse.	Same.

Table 2.1: *Description of a target trial to identify the optimal androgen deprivation therapy timing with respect to prostate-specific antigen doubling time, and of its emulation in observational data. (continued)*

Components	Target trial	Emulation of trial
Eligibility criteria	1. Histologically confirmed adenocarcinoma of the prostate, clinically staged cT3aN0M0 or lower.	1. Same.
	2. PSA relapse after definitive radical treatment (prostatectomy and/or radiotherapy), as evidenced by one of the following: a) PSA rise above 0.2 ng/mL beyond post-treatment nadir if initial treatment was prostatectomy (with or without radiation); b) PSA that did not fall lower than 0.2 ng/mL if treated with prostatectomy with salvage radiation; c) Three successive PSA rises at least 30 days apart if initial treatment was only radiation.	2. Same. 3. Same. Operationalized as lack of any a) Positive findings on pelvis MRI, abdomen CT, pelvis CT, or bone scan at any time in the past; b) Severe symptoms (fatigue, bone pain, anorexia, weight loss, abdominal pain) at the time of relapse; c) Progression noted in physician notes.
	3. No symptomatic disease requiring therapy, or any evidence of metastatic disease.	4. Same. Operationalized as: a) Never had orchiectomy; b) No prescribed ADT within the past year, or for more than 12 months at any time.
	4. Naive to ADT treatment (no orchiectomy, and any previous ADT was more than 1 year in the past, and not for longer than 12 months).	5. Same. Operationalized based on National Comprehensive Cancer Network guidelines principles of life expectancy estimation, using the Social Security Administration life tables.
	5. Life expectancy at least 5 years.	6. Same.
	6. PSA doubling time at relapse of 30 days or more.	

Table 2.1: *Description of a target trial to identify the optimal androgen deprivation therapy timing with respect to prostate-specific antigen doubling time, and of its emulation in observational data. (continued)*

Components	Target trial	Emulation of trial
Treatment strategies	Initiate ADT within 3 months after PSA doubling time drops below a prespecified value, from 0 to 360 days in increments of 10. Under all strategies, ADT will be started within 3 months after a patient experiences further disease progression. Continuation of ADT after initiation will be left at the physician's and patient's discretion. Once ADT has been discontinued, it will not be reinitiated. Under all strategies, individuals will continue to have PSA measured and symptoms assessed at the physician's and patient's discretion. Patients complete study-specific surveys at baseline and every six months thereafter.	Same.
Treatment assignment	Each individual is randomized a treatment strategy defined by a PSADT threshold.	Each individual is assigned to all treatment strategies.
Follow-up	Patients are followed from treatment assignment (time zero) until death, loss to follow-up (24 months without contact with the study team via returned surveys, or physician visits), or administrative end of follow-up.	Same.
Outcome	Death from any cause.	Same.
Causal contrast	Intention-to-treat effect and per-protocol effect.	Observational analog of per-protocol effect.

Table 2.1: *Description of a target trial to identify the optimal androgen deprivation therapy timing with respect to prostate-specific antigen doubling time, and of its emulation in observational data. (continued)*

Components	Target trial	Emulation of trial
Analysis plan	Intention-to-treat analysis: Survival curves and 5- and 10-year mortality estimates within each treatment arm. Adjustment for potential selection bias due to loss to follow-up via IP weighting. Per-protocol analysis: same except that individuals are censored at non-adherence and uncensored individuals are assigned IP weighs that are a function of baseline and time-varying variables. Alternatively, the per-protocol analysis may be based on the g-formula.	Observational analog of a per-protocol analysis: same as in target trial, except that analyses are not conducted separately by assigned treatment strategy, and we created 37 individuals (clones) per eligible patient and assigned one to each strategy when using censoring plus IP weighting.

Abbreviations
PSA, prostate-specific antigen; ADT, androgen deprivation therapy; PSADT, prostate-specific antigen doubling time.

Target trial: Intention-to-treat analysis

In a small abuse of notation, we refer to strategy x as the strategy in which treatment is initiated within 3 months after PSADT drops below x or disease progresses. To estimate the intention-to-treat effect, we compare the survival curves between individuals assigned to each strategy x. That is, we estimate $\Pr(Y_t = 1 \mid X = x)$ where $t = 0, \ldots, 120$ months of follow-up and Y_t an indicator of death from any cause during or before month t.

We could estimate $\Pr(Y_t = 1 \mid X = x)$ nonparametrically. However, with so many treatment arms, we may wish to obtain more precise estimates by making parametric assumptions, e.g., by fitting a pooled logistic regression model for the discrete-time hazard $\Pr(Y_t = 1 \mid X = x, Y_{t-1} = 0)$ with a time-varying intercept, modeled as natural cubic spline terms, and a covariate for treatment

strategy x, also modeled as cubic splines and product ("interaction") terms with time. The model's predicted values are then used to compute the survival curve for each strategy.[54]

Additionally, if imbalances existed in baseline characteristics across groups, the model would include them as covariates. We would then standardize the estimated probabilities to the distribution of the covariates to estimate marginal survival curves. Finally, if necessary, inverse probability (IP) weighting would be used to adjust for selection bias from loss to follow-up.[63]

Target trial: Per-protocol analysis

To estimate the per-protocol effect, we would compare the survival curves under adherence to each of the strategies. Because adherence is not randomized, we would need appropriate adjustment for (possibly time-varying) confounders, that is, prognostic factors that are determinants of adherence to the assigned treatment strategy, or their proxies. Let L_t be the vector of measured covariates in month t, including an indicator of having a clinic visit in month t, an indicator of disease progression, an indicator of symptoms (bone pain, fatigue, weight loss, anorexia, and abdominal pain), and PSA. L_0 contains baseline covariates: D'Amico risk group, comorbidities, and age at diagnosis; time from diagnosis to relapse; and calendar year, PSA, and PSADT at relapse.

To adjust for these confounders, we can use IP weighting or the parametric g-formula. Both methods have been described elsewhere;[54,56] we review them here. Under the assumption that losses to follow-up and non-adherence happen at random within levels of the confounders, and that all models (described below) are correctly specified, both methods consistently estimate the survival probabilities had everyone adhered to each strategy and stayed under follow-up throughout the duration of the study. We use the non-parametric bootstrap with 1000 samples to estimate 95% confidence intervals under each approach.

IP weighting of a dynamic marginal structural model

We fit the same pooled logistic model as for the intention-to-treat analysis with two modifications.

First, individuals are censored if/when their observed treatment and covariate history is no longer consistent with their assigned strategy x. Censoring can occur for three reasons: the individual initiates treatment before their PSADT drops below x or before they experience disease progression; the individual does not initiate treatment within the 3-month grace period after meeting either of the criteria to start treatment; or the individual begins treatment again after having previously concluded it. During the grace period, no individual can be censored. Also, according to the protocol, individuals can stop at any time after treatment initiation, so no individuals can be censored while receiving treatment. However, the protocol does not allow for treatment re-initiation after discontinuation, so individuals will be censored if they begin treatment after discontinuing it.

Second, to adjust for the potential selection bias introduced by censoring for non-adherence, at each month t we assign a time-varying IP weight to each individual. The denominator of the weight is the probability of remaining uncensored through month t, which is equal to the probability of remaining uncensored in months $k = 0, ..., t$. The probability of being uncensored is $f(A_k \mid \bar{L}_k, \bar{A}_{k-1}, \overline{Y}_{k-1} = 0)$, where A_t is an indicator of treatment during month t and $\bar{A}_t$ the treatment history from time 0 through time t, during months with $A_{k-1} = 0$. Because our treatment strategies allow treatment discontinuation at any time, the denominator is 1 for months with $A_{k-1} = 1$. We can estimate $f(A_k \mid \bar{L}_k, A_{k-1} = 0, \bar{A}_{k-2}, \overline{Y}_{k-1} = 0)$ via a pooled logistic model, separately within treatment arms and separately among months with and without prior use of treatment during the study.

If the protocol of the target trial specifies strategies of the form: "Start treatment *with equal probability* during any of the three months following the [initiation threshold]," then we will ensure that the per-protocol effect is estimated under this initiation pattern by multiplying the weights during the grace period by an additional factor.[54] For a grace period of 3 months, that

factor is $\frac{1}{4}$ for an initiator in the first month of eligibility, $\frac{1}{3}$ in the second, and $\frac{1}{2}$ in the third; for a non-initiator, the factors are $\frac{3}{4}$, $\frac{2}{3}$, and $\frac{1}{2}$, respectively.

Finally, individuals are also censored at the end of any two-year period in which they did not visit a physician or complete a survey at least once. We can also estimate IP weights to adjust for potential selection bias due to this censoring.[6]3

Parametric g-formula

The g-formula can be viewed as a generalized form of standardization of the conditional hazard under each treatment strategy to the joint distribution of the time-varying covariates. To estimate each component of the g-formula, we can fit within each treatment arm a logistic model for $\Pr(Y_{t+1} = 1 \mid \bar{L}_t, \bar{A}_t, Y_t = 0)$, and logistic or linear models for the conditional density of each of the time-varying covariates in the vector L_t. We also need to fit a logistic model for the conditional probability of discontinuation of treatment A_t because the protocol does not prescribe the probability of stopping treatment after initiating. In contrast, the probability of treatment initiation under each strategy is known: 0 before reaching the PSADT threshold, 1 within 3 months of reaching it ($\frac{1}{4}$ in the first month, $\frac{1}{3}$ in the second, $\frac{1}{2}$ in the third, and 1 at the end of the grace period), and 0 again if treatment is discontinued.

Finally, we need to assign a monitoring strategy that aligns with our trial protocol. The probability of a clinic visit is estimated using a model in the observed data, but is set to 1 if there has been no visit in the last 23 months, guaranteeing monitoring at least every 2 years.

We then standardize the probability of the outcome under each strategy x by averaging over all treatment and covariate histories, using the modeled densities. The resulting integral can be approximated via Monte Carlo simulation.

Target trial: Emulation

The Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE) observational study of prostate cancer patients began enrollment in 1995. The study has been described in detail elsewhere.[65] We used data through 2017, at which point over 14,000 biopsy-proven patients had been enrolled at over 40 U.S. clinics and followed prospectively. Physicians provided clinical data (diagnosis, start and stop dates of medications, outcomes, lab and imaging tests) and participants provided a follow-up survey approximately every 6-12 months after a baseline questionnaire (quality of life, other health-service use). Evidence of disease progression after relapse was based on clinical notes describing severe symptoms or metastases seen on imaging. We used consecutive PSA measurements to calculate PSADT. We set to the value of the 95th percentile of PSADT those values that were greater than the 95th percentile (because of very slow-growing PSA)[62] or were undefined because PSA was constant or decreased from one date to the next.[66]

We used these data to emulate the eligibility criteria, treatment strategies, outcome, and follow-up of the target trial as summarized in Table 2.1.

Per-protocol analysis via IP weighting of a dynamic marginal structural model

We carried out the analysis described for the target trial (specifications for each model are shown in Table B.2). However, since treatment strategies are not assigned at random in observational studies, some modifications had to be made. First, we estimated the probability of treatment initiation among all eligible individuals, instead of separately within treatment arms. Second, because there was no assignment of each individual to a single strategy, we allowed for each individual to be part of the analysis for *each* strategy by copying the dataset for each value of x considered, and then censoring participants separately within each dataset when their observed data were not consistent with that strategy.[53,54] We truncated the total weights at the 99th percentile to avoid near positivity violations.

Per-protocol analysis via the parametric g-formula

The estimation procedures for the observational emulation were the same as for the target trial. We used the R package `gfoRmula`.[67] All analyses were conducted in R version 3.4.3.[68]

Sensitivity analyses

To explore the effects of our choice of target trial protocol, we repeated the two analyses using a different distribution for treatment initiation during the grace period. We specified that the rate of treatment initiation during the grace period would be the same that would have been observed in the absence of an intervention until the end of the grace period, at which point treatment would be initiated if it had not been previously. For the IP weighting approach, this meant that the factors in the weights for treatment during the grace period were equal to 1. For the g-formula, we additionally fit a model for treatment distribution and during the grace period drew treatment values with probabilities estimated from that model.

In addition, we investigated whether assigning treatment initiation thresholds based on another function of PSA would better target those in need of treatment. Specifically, we repeated our original analyses but assigned treatment when *average* PSADT since relapse reached cutoffs between 0 and 1800, in increments of 150.

Finally, we conducted an unadjusted analysis by fitting an unweighted pooled logistic regression model for mortality in the censored and concatenated datasets, using only the terms for time and treatment strategy.

2.3 Results

After applying the eligibility criteria, we found 1,229 eligible individuals (Figure 2.1). Their baseline characteristics are shown in Table 2.2. About 60% underwent radical prostatectomy as their

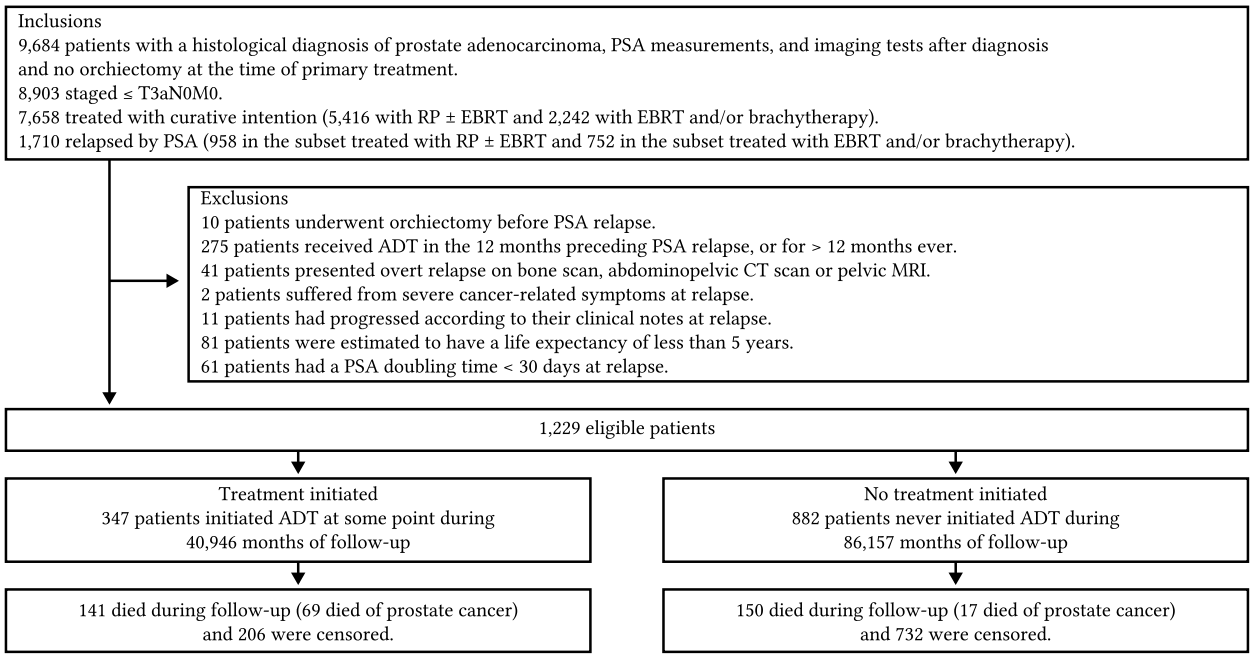

Figure 2.1: *Flowchart of patient selection from the Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE) database through 2017 into the present study.*

original treatment, and 47% were assigned to a medium clinical risk group at that time. The median time to biochemical recurrence after diagnosis was 3.3 years.

Table 2.2: *Baseline characteristics of the analytic sample of prostate cancer patients with biochemical recurrence in the Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE), 1995-2017 (n = 1229).*

Characteristic	n (%)
Vital status	
Alive at end of follow-up	972 (79%)
Died during follow-up	257 (21%)
Relapse date	
1980s	2 (0.2%)
1990s	406 (33%)
2000s	741 (60%)
2010s	80 (6.5%)
Years to relapse[a]	3.29 (2.10, 5.35)
Original treatment	
Radical prostatectomy	733 (60%)

Table 2.2: *Baseline characteristics of the analytic sample of prostate cancer patients with biochemical recurrence in the Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE), 1995-2017 (n = 1229). (continued)*

Characteristic	n (%)
Radiotherapy	496 (40%)
PSA at relapse[a]	0.52 (0.32, 1.40)
PSADT at relapse (days)[a]	256 (124, 616)
ADT after relapse	345 (28%)
Comorbidities	
0 or 1	512 (42%)
More than 1	504 (41%)
Missing	213 (17%)
Clinical risk group at diagnosis	
Low	397 (32%)
Medium	576 (47%)
High	256 (21%)
Age at diagnosis	
40-49	22 (1.8%)
50-59	246 (20%)
60-69	577 (47%)
70-79	367 (30%)
80-89	17 (1.4%)

[a] Median (interquartile range)

Of the participants, 347 actually received androgen deprivation therapy of any kind at some point during follow-up, and 291 died from any cause. Because many individuals never initiated treatment, there were fewer person-months in the data that were consistent with treatment strategies defined by higher PSADT thresholds. For the treatment strategy defined by a threshold of 0, there were 64,247 person-months and 145 deaths. For the treatment strategy defined by a threshold of 360, there were 25,205 person-months and 64 deaths (Table B.1).

The estimated survival was similar under all strategies (Figure 2.2), though estimates were imprecise. Risk differences for 10-year mortality comparing the highest threshold (360) with the lowest (0) (i.e., earlier vs. later initiation) were 0.02 (-0.31, 0.44) and -0.02 (-0.05, 0.04) when

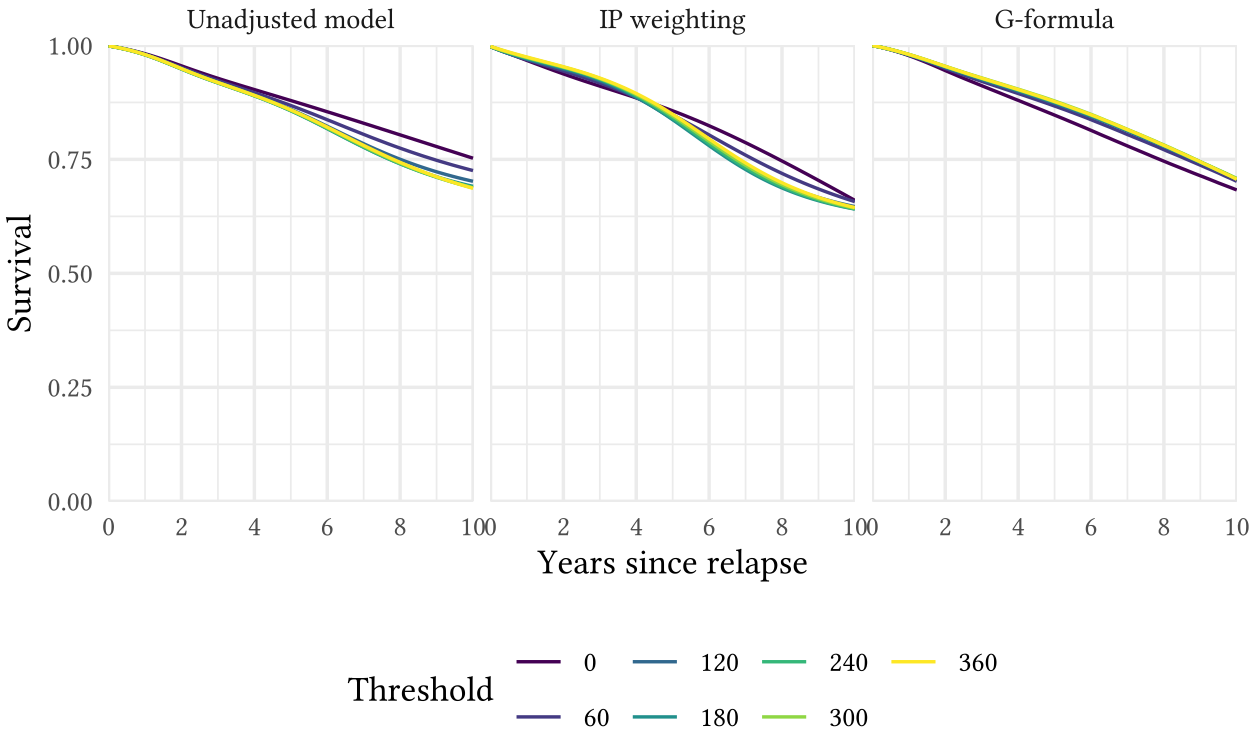

Figure 2.2: *Survival curves estimated via various methods, comparing treatment strategies defined by PSA doubling time thresholds. Thresholds range from* 0 *(darkest purple, least treatment) to* 360 *(brightest yellow, most treatment).*

estimated via IP weighting and with the parametric g-formula, respectively (Table B.3). Results were similar when we varied the target trial protocol and the PSADT truncation level (Table B.3), and the treatment thresholds (Table B.4).

2.4 Discussion

We estimated the per-protocol effect of treatment strategies for prostate cancer using observational data. We showed that the treatment strategy needs to be unambiguously specified by describing valid reasons for treatment initiation, as well the grace period for initiation and the patterns of initiation during the grace period. Though some of these components may be unacknowledged when estimating the per-protocol effect via IP weighting of a dynamic marginal structural model, their specification is required when using the parametric g-formula.

Treatment after initiation requires the same precise specification. Our treatment protocol excluded intermittent treatment strategies, instead requiring that patients initiate treatment only once. However, intermittent treatment with androgen deprivation therapy, in which treatment is stopped and reintroduced multiple times based on closely monitored PSA and testosterone levels, has similar efficacy with respect to overall survival as continuous therapy, with possibly better quality of life.[69] When complex rules may define treatment continuation after initiation, a protocol in which such treatment is assigned to follow patterns actually observed in the data is simplest with the IP weighting approach: censoring after treatment initiation only occurs due to lack of monitoring or loss to follow-up, and not for changes in treatment. Such a treatment protocol would allow for both continuous and intermittent strategies at the same relative frequency in which they occurred at the clinics from which the data was collected.

However, this approach is not easily extensible to the g-formula, which requires explicitly specifying those treatment patterns. While we allowed the duration of the first bout of treatment to reflect the observed data – in the IP weighting approach by not censoring anyone for continuing or ceasing treatment, and with the g-formula by modeling the monthly probability of treatment discontinuation in the data and accordingly assigning discontinuation – allowing for additional treatment as observed in the data would have required more complex modeling. Alternatively, a protocol that explicitly defined a single intermittent treatment strategy for everyone would have been easier to implement with the g-formula, but would have resulted in censoring many more observations with the IP weighting approach. Given the geographic and temporal variability in our data, few participants are likely to have followed any one strategy.

Among the treatment strategies we did compare, wide 95% confidence intervals imply that our data are equally compatible with harm, benefit, or no effect of early initiation of androgen deprivation therapy on survival. As expected, confidence intervals from the parametric g-formula approach were narrower, reflecting the additional parametric assumptions compared with the IP weighting approach.

Little prior evidence is available to determine the optimal treatment in prostate cancer patients with asymptomatic biochemical recurrence, and this study does not add any conclusive evidence. Ideally, risk of deadly metastatic disease should be balanced against the threat to quality of life that hormonal treatment poses. US guidelines refrain from recommending a standard treatment in this situation due to uncertainty about timing after early signs of biochemical recurrence.[70] One reason for hesitation in assigning all relapsing patients to immediate therapy is the prolonged timeline of cancer spread in most patients. On average, clinical metastasis becomes apparent 7-8 years after biochemical recurrence; due to the age of the affected population, this is around or beyond many patients' expected lifespan even without cancer.[50] One trial investigating delayed treatment found that 41% of individuals in the delayed therapy arm died before reaching the point at which they would initiate androgen deprivation therapy.[58] Furthermore, the therapy leads to a number of side effects, including weight gain and loss of muscle mass, osteoporosis and anemia, and sexual dysfunction, which can reduce quality of life with possibly little benefit.

In conclusion, we found little evidence that initiating androgen deprivation therapy on the basis of PSA doubling time reduces all-cause mortality. Estimating counterfactual quantities under exactly the same strategy with both the weighting and g-formula approaches allows for direct comparison of the approaches. Since we aligned the study protocols, differences reflect random error and possible model misspecification, but not different estimands. However, while comparable results are reassuring, both methods rely on the same measured confounders and are subject to violations of the identification assumptions. We made some simplifying assumptions about the time-varying confounders that may have resulted in residual confounding. In addition, there may be unmeasured confounding if the decision to initiate treatment was made on the basis of factors not recorded by the clinician or participant, or not included in our analysis. We made similar assumptions about the predictors of loss to follow-up; informative censoring could also have biased our results.

When the target trial is a real randomized study that has been completed or that will be conducted in the future, the results of an observational emulation can be compared to those

from the trial to investigate any discrepancies.[71] A well-conducted emulation (e.g., with aligned eligibility criteria, treatment and outcome definitions, length of follow-up) reduces the possibility of such discrepancies.

3

Multiple-bias sensitivity analysis using bounds

Confounding, selection bias, and measurement error are well-known sources of bias in epidemiologic research. Methods for assessing these biases have their own limitations. Many quantitative sensitivity analysis approaches consider each type of bias individually, while more complex approaches are harder to implement or require numerous assumptions. By failing to consider multiple biases at once, researchers can underestimate – or overestimate – their joint impact. We show that it is possible to bound the *total* composite bias due to these three sources, and to use that bound to assess the sensitivity of a risk ratio to any combination of these biases. We derive bounds for the total composite bias under a variety of scenarios, providing researchers with tools to assess their total potential impact. We apply this technique to a study where unmeasured confounding and selection bias are both concerns, and to another study in which possible differential exposure misclassification and confounding are concerns. The approach we describe, though conservative, is easier to implement and makes simpler assumptions than quantitative bias analysis. We provide R functions to aid implementation.

This chapter was co-authored with Maya B. Mathur and Tyler J. VanderWeele.

3.1 Introduction

Assessing evidence for causation is fundamental in order to plan and target interventions and improve public health. However, many causal claims in epidemiologic studies are met with suspicion by both researchers and the general public, due to the fact that such studies are well known to be subject to various biases. While faults in these studies can sometimes be addressed directly – e.g., through better sampling schemes, blinded outcome ascertainment, more extensive covariate measurements, etc. – other times confounding, selection bias, and measurement error are unavoidable. In such situations, our next best option is to assess the extent to which a given study's conclusions might be sensitive or robust to these biases, and whether they threaten its conclusions. Often, however, this is limited to a few sentences in a discussion section qualitatively assessing the possibility of bias, sometimes appealing without quantitative justification to heuristics that may or may not hold true in a particular study.[72–74]

The weak uptake of quantitative bias analysis in epidemiology belies its long history. Over a half-century ago, Cornfield and then Bross argued that the extent of possible bias was quantifiable based on observed data and possibly hypothetical quantities.[75–78] Attempts to generalize these results, as well as consider other biases, sometimes simultaneously with confounding, followed.[78–83] More recently, probabilistic bias analysis methods have been developed, allowing researchers to propose distributions for various bias parameters across multiple biases, and to explore how various combinations of those parameters would affect their results.[84–89] Despite the availability of these methods in textbook and software form,[87,90,91] the actual uptake of such quantitative bias analysis in empirical research has been limited,[92] possibly because of the (at least perceived) computational complexity[93] or difficulty in proposing plausible distributions.

Even more recently, simpler approaches to sensitivity analysis have hearkened back to the early days of bias assessment, with the development of bounds for various biases that require limited assumptions and at most basic algebra.[94–96] However, a one-at-a-time approach is not sufficient for many studies subject to multiple sources of bias. In this article we extend the simple

sensitivity analysis framework to multiple biases, describing a bound for the total composite bias from confounding, selection, and differential exposure or outcome misclassification.

3.2 The problem of multiple biases

We will describe a scenario in which all three types of bias are present, preventing the interpretation of an observed risk ratio as a causal risk ratio. Consider a binary exposure A, a binary outcome Y, and measured covariates C. (C may also include unmeasured factors that are controlled by the study design.) Let S be an indicator of the subset of the population for which data have been collected, and let A^* and Y^* denote misclassified versions of the exposure and outcome, respectively. We use potential outcome notation to describe causal quantities: Y_a is the outcome that would occur were exposure A set to value a. We assume consistency, meaning that $Y_a = Y$ for observations for whom we observe $A = a$, and positivity, meaning that $0 < \Pr(A = 1 \mid \cdot) < 1$ within every stratum of the population.

We denote (conditional) independence between random variables with the symbol $\perp\!\!\!\perp$, such that $Y_a \perp\!\!\!\perp A \mid C$ implies conditional exchangeability; i.e., potential outcomes are independent of exposure status conditional on C. However, when C does not capture all of the exposure-outcome confounding, it is not true that $Y_a \perp\!\!\!\perp A \mid C$. We assume in that case that additionally adjusting for some unmeasured factor(s) U_c would be sufficient to address confounding, so that $Y_a \perp\!\!\!\perp A \mid C, U_c$. Here, U_c may be a single random variable or a vector of variables, which may be continuous or take on any number of discrete values, or some combination. Similarly, we allow for selection bias, which we define as a lack of the conditional independence $Y_a \perp\!\!\!\perp A \mid C, U_c, S = 1$ when it is otherwise true that $Y_a \perp\!\!\!\perp A \mid C, U_c$. We likewise assume that the measurement of some variable(s) U_s, responsible for selection, would fully account for this bias, though the necessary conditions for it to do so will depend on whether we intend to make inferences about effects in the total population, or just the selected population. Finally, we allow for the possibility that the misclassification is differential, by which we mean that the sensitivity or specificity of the exposure measurement may

differ depending on the value of the outcome, or that the sensitivity or specificity of the outcome measurement may depend on the exposure. In our notation, this means that it is not necessarily true that $A^* \perp\!\!\!\perp Y \mid A, C$ or that $Y^* \perp\!\!\!\perp A \mid Y, C$. In this work we consider only misclassification of the exposure or the outcome, but not both at once.

Motivating examples

There is great interest in how exposures during pregnancy may affect offspring health. However important such questions are, they are difficult to answer with epidemiologic research. Ethics may limit inclusion of pregnant people in randomized trials, and many exposures of interest are not ethical or feasible to randomize to anyone. Case-control studies can efficiently capture rare childhood outcomes, but recalling pregnancy exposures several years later can result in measurement error.[97] Prospective cohort studies can avoid this recall bias, but are often subject to loss to follow-up when the duration between exposure and outcome assessment is long.[98] Observational studies of all types are threatened by uncontrolled confounding, and inter-generational confounders are particularly difficult to assess.[99] Importantly, studies like these are not affected by only one or another of these biases, but may suffer from multiple threats to validity.

To demonstrate our sensitivity analysis approach, we will consider two questions about exposures during pregnancy and outcomes in children: whether HIV infection in utero causes wasting (low weight-for-length), and whether vitamin consumption during pregnancy protects against childhood leukemia.

Omoni et al. investigated the former hypothesis concerning HIV infection and wasting (among participants of a vitamin A supplementation trial in Zimbabwe) and found that, compared to children who were unexposed to HIV, those who had been infected with HIV in utero were more likely to be below a weight-for-length Z-score of -2 as toddlers.[100] The odds ratio comparing the two groups was 6.75 (95% CI, 2.79, 16.31) at 2 years. Although randomized trial data were used for the analysis, this was an observational study with respect to HIV infection, since infection is

not randomized. The authors did not, however, adjust for any confounders. Furthermore, since enrollment occurred at delivery, after possible HIV exposure and transmission, the choice of whether to participate could have been affected by HIV status as well as other factors, leading to selection bias if those factors affect future child growth. We will consider the role that confounding and selection bias may play in this study.

As a second example, Ross et al. analyzed the relationship between vitamins and leukemia in a case-control study and found a decreased risk of acute lymphoblastic leukemia among children whose mothers consumed vitamin supplements during pregnancy.[101] Their reported odds ratio, which, with a rare outcome, approximates a risk ratio, of 0.51 (95% CI 0.30, 0.89) was conditional on maternal age, race, and a binary indicator of education. However, there may be other confounders that were not controlled, such as other indicators of a privileged or healthy lifestyle that are both associated with vitamin use and protective against leukemia. We also may be concerned about recall bias (differential exposure misclassification) – that mothers of children with a cancer diagnosis might be more likely to report *not* taking a vitamin even if they did so – so we consider how exposure misclassification and unmeasured confounding can be assessed simultaneously.

3.3 The multiple-bias bound

Two overarching types of bias analysis have been described: one that explores how biases of a given magnitude affect an estimate, which Phillips labeled "bias-level sensitivity analysis," and another that reduces the analysis to a summary of how much bias would be necessary for an observation to be compatible with a truly null effect (or some other specified non-null effect), which he called "target-adjusted sensitivity analysis."[73] We focus here on the former and address the latter in the eAppendix. Here we present a multiple-bias bound, which allows researchers or consumers of research to explore the maximum factor by which unmeasured confounding, selection, and misclassification could bias a risk ratio.

We begin with outcome misclassification, and then extend our results to exposure misclassification. We assume that the investigators have estimated $\mathrm{RR}^{\mathrm{obs}}_{AY^*} = \frac{\Pr(Y^*=1|A=1,S=1,c)}{\Pr(Y^*=1|A=0,S=1,c)}$, the observed risk ratio conditional on some value c of the covariates, but wish to make inference about the causal conditional risk ratio $\mathrm{RR}^{\mathrm{true}}_{AY} = \frac{\Pr(Y_1=1|c)}{\Pr(Y_0=1|c)}$. We will assess bias on the relative scale, so that we define the bias as $\mathrm{RR}^{\mathrm{obs}}_{AY^*}/\mathrm{RR}^{\mathrm{true}}_{AY}$.

Using bounds that have been previously described for misclassification, selection bias, and unmeasured confounding considered individually,[94–96,102] we can bound $\mathrm{RR}^{\mathrm{obs}}_{AY^*}$ by factoring it into $\mathrm{RR}^{\mathrm{true}}_{AY}$ and components for each of the biases. The parameters that will be used to bound the biases are as follows:

$$\mathrm{RR}_{AY^*|y,S=1} = \max_y \frac{\Pr\left(Y^*=1 \mid Y=y, A=1, S=1, c\right)}{\Pr\left(Y^*=1 \mid Y=y, A=0, S=1, c\right)}$$

$$\mathrm{RR}_{U_sY|A=a} = \frac{\max_u \Pr\left(Y=1 \mid A=a, c, U_s=u\right)}{\min_u \Pr\left(Y=1 \mid A=a, c, U_s=u\right)} \quad \text{for } a=0,1$$

$$\mathrm{RR}_{SU_s|A=a} = \max_u \frac{\Pr\left(U_s=u \mid A=a, S=a, c\right)}{\Pr\left(U_s=u \mid A=a, S=1-a, c\right)} \quad \text{for } a=0,1$$

$$\mathrm{RR}_{U_cY} = \max_a \frac{\max_u \Pr\left(Y=1 \mid A=a, c, U_c=u\right)}{\min_u \Pr\left(Y=1 \mid A=a, c, U_c=u\right)}$$

$$\mathrm{RR}_{AU_c} = \max_u \frac{\Pr\left(U_c=u \mid A=1, c\right)}{\Pr\left(U_c=u \mid A=0, c\right)} \quad .$$

These bias parameters have been described elsewhere, though separately.[94–96] Briefly, the bias parameter defining the misclassification portion of the bound ($\mathrm{RR}_{AY^*|y,S=1}$) describes the maximum of the false positive probability ratio or sensitivity ratio *within* the selected population. The selection bias parameters ($\mathrm{RR}_{U_sY|A=a}$ and $\mathrm{RR}_{SU_s|A=a}$) describe the maximum factors by which the outcome risk differs by values of U_s, within strata of A, and the maximum factors by which some level of U_s differs between the selected and non-selected groups, within strata of A. Finally, the unmeasured confounding parameters (RR_{U_cY} and RR_{AU_c}) describe the maximum factor by which U_c increases the outcome risk, conditional on A, and the maximum factor by which exposure is associated with some value of U_c. Each of the sensitivity parameters is conditional on the

covariates adjusted for in the analysis, and so describes the extent of bias above and beyond those factors.

To simplify notation, define the function $g(a,b) = \frac{a \times b}{a+b-1}$. Then we have the following bound for the total composite bias.

Result 1:

If $Y_a \perp\!\!\!\perp A \mid C, U_c$ and $Y \perp\!\!\!\perp S \mid A, C, U_s$, then:

$$\mathrm{RR}_{AY^*}^{\mathrm{obs}} / \mathrm{RR}_{AY}^{\mathrm{true}} \leq \mathrm{BF}_m \times \mathrm{BF}_s \times \mathrm{BF}_c$$

where $\mathrm{BF}_m = \mathrm{RR}_{AY^*|y,S=1}$, $\mathrm{BF}_s = g\left(\mathrm{RR}_{U_s Y|A=1}, \mathrm{RR}_{SU_s|A=1}\right) \times g\left(\mathrm{RR}_{U_s Y|A=0}, \mathrm{RR}_{SU_s|A=0}\right)$, and $\mathrm{BF}_c = g\left(\mathrm{RR}_{AU_c}, \mathrm{RR}_{U_c Y}\right)$. The derivation of this and the results that follow are given in the eAppendix.

Result 1 can be used to quantify the maximum amount of bias that could be produced by parameters of a given value. Values for the sensitivity parameters may be taken from validation studies, previous literature, or expert knowledge, or proposed as hypotheticals. Because the sensitivity parameters are maxima, they are always greater than or equal to 1 and the composite bound will thus be greater than or equal to 1. For an apparently causative observed exposure-outcome risk ratio (>1) one could divide the estimate and its confidence interval by the bound to obtain the maximum that the specified biases could shift the estimate and its confidence interval. For a preventive observed exposure-outcome risk ratio (<1) one could multiply the estimate and its confidence interval by the bound to obtain the maximum that the specified biases could shift the estimate and its confidence interval, or equivalently reverse the coding of the exposure to obtain a risk ratio >1. By applying the bound to the confidence interval closet to the null, we can make statements such as: "In 95% of repeated samples with the same sources of bias, adjusting the confidence interval in this way would result in a lower bound that is less than the true causal risk ratio, provided the proposed parameter values adequately bound (i.e., are as large as or larger than) the true parameter values."

Although the bound allows for terms for all three biases, if any of them is judged not to threaten a given study, or to bias toward the null, that factor can be omitted. Furthermore, the

selection bias term can be simplified under certain assumptions;[95] we illustrate in the first example below.

Result 1 is summarized in the first row of Table 3.1. The assumptions required for the bound to hold are listed under the biases to which they pertain, and the bound itself is in the final column. Note that the factorization of the bound implies an ordering of the biases: the misclassification parameters are defined within the stratum $S = 1$. Intuitively, this corresponds to a study in which outcome measurement is done after people have been selected into the study, and so requires considering the strength of differential misclassification only within that group. In general, we can think of biases as layers that we must peel off sequentially, and the order in which we do so is the reverse of the order in which they occurred in the data.[103,104] Confounding is generally thought of as a property of nature within the population of interest, so occurs first (though if parameters describing the strength of confounding are derived based on misclassified exposure or outcome, that may not be the case[103]), but the order in which selection and misclassification occur may depend on the study design. We could alternatively derive a bound that depends on a parameter describing the extent to which the outcome is misclassified in the total population, and on others describing how selection is associated with the misclassified outcome. These parameters may be more intuitive in a study with case-control sampling. Additionally, another ordering of the biases may be preferable if data exists to justify estimates of the alternative parameters. We define those alternative parameters and derive that bound in the eAppendix. The second row of Table 3.1 summarizes those results; there, the assumption for selection bias is an assumption about the misclassified outcome, and the bound in the final column is defined in terms of parameters that reflect that ordering.

Table 3.1: *Multiple bias bounds for various combinations of biases. The first three columns show different combinations and ordering of the biases, as well as the implied assumptions. The fourth column contains the expression that bounds the bias when those assumptions hold. The definitions of the parameters are given in the main text.*

Biases and associated assumptions			Bound under the stated assumptions
Bias 1	Bias 2	Bias 3	
unmeasured confounding $Y_a \perp\!\!\!\perp A \mid C, U_c$	general selection bias $Y \perp\!\!\!\perp S \mid A, C, U_s$	outcome misclassification	$g\left(\mathrm{RR}_{AU_c}, \mathrm{RR}_{U_cY}\right) \times g\left(\mathrm{RR}_{U_sY\mid A=1}, \mathrm{RR}_{SU_s\mid A=1}\right) \times$ $g\left(\mathrm{RR}_{U_sY\mid A=0}\mathrm{RR}_{SU_s\mid A=0}\right) \times \mathrm{RR}_{AY^*\mid y,S=1}$
unmeasured confounding $Y_a \perp\!\!\!\perp A \mid C, U_c$	outcome misclassification	general selection bias $Y^* \perp\!\!\!\perp S \mid A, C, U_s$	$g\left(\mathrm{RR}_{AU_c}, \mathrm{RR}_{U_cY}\right) \times \mathrm{RR}_{AY^*\mid y} \times$ $g\left(\mathrm{RR}_{U_sY^*\mid A=1}, \mathrm{RR}_{SU_s\mid A=1}\right) \times g\left(\mathrm{RR}_{U_sY^*\mid A=0}\mathrm{RR}_{SU_s\mid A=0}\right)$
unmeasured confounding $Y_a \perp\!\!\!\perp A \mid S=1, C, U_c, U_s$	selected population	outcome misclassification	$g\left(\mathrm{RR}_{AU_{sc}\mid S=1}, \mathrm{RR}_{U_{sc}Y\mid S=1}\right) \times \mathrm{RR}_{AY^*\mid y}$
unmeasured confounding $Y_a \perp\!\!\!\perp A \mid C, U_c$	general selection bias $Y \perp\!\!\!\perp S \mid A, C, U_s$	exposure misclassification $\Pr(Y=0 \mid a, c, S=1) \approx 1$	$g\left(\mathrm{RR}_{AU_c}, \mathrm{RR}_{U_cY}\right) \times g\left(\mathrm{RR}_{U_sY\mid A=1}, \mathrm{RR}_{SU_s\mid A=1}\right) \times$ $g\left(\mathrm{RR}_{U_sY\mid A=0}\mathrm{RR}_{SU_s\mid A=0}\right) \times \mathrm{OR}_{YA^*\mid a,S=1}$
unmeasured confounding $Y_a \perp\!\!\!\perp A \mid C, U_c$	exposure misclassification $\Pr(Y=0 \mid a, c) \approx 1$	general selection bias $Y \perp\!\!\!\perp S \mid A^*, C, U_s$	$g\left(\mathrm{RR}_{AU_c}, \mathrm{RR}_{U_cY}\right) \times \mathrm{OR}_{YA^*\mid a} \times$ $g\left(\mathrm{RR}_{U_sY\mid A^*=1}, \mathrm{RR}_{SU_s\mid A^*=1}\right) \times$ $g\left(\mathrm{RR}_{U_sY\mid A^*=0}\mathrm{RR}_{SU_s\mid A^*=0}\right)$
unmeasured confounding $Y_a \perp\!\!\!\perp A \mid S=1, C, U_c, U_s$	selected population	exposure misclassification $\Pr(Y=0 \mid a, c, S=1) \approx 1$	$g\left(\mathrm{RR}_{AU_{sc}\mid S=1}, \mathrm{RR}_{U_{sc}Y\mid S=1}\right) \times \mathrm{OR}_{YA^*\mid a}$
unmeasured confounding $Y_a \perp\!\!\!\perp A \mid C, U_c$	general selection bias $Y \perp\!\!\!\perp S \mid A, C, U_s$	exposure misclassification $\Pr(Y=0 \mid a, c, S=1) \approx 1$ $\Pr(A=0 \mid y, c, S=1) \approx 1$	$g\left(\mathrm{RR}_{AU_c}, \mathrm{RR}_{U_cY}\right) \times g\left(\mathrm{RR}_{U_sY\mid A=1}, \mathrm{RR}_{SU_s\mid A=1}\right) \times$ $g\left(\mathrm{RR}_{U_sY\mid A=0}\mathrm{RR}_{SU_s\mid A=0}\right) \times \mathrm{RR}_{YA^*\mid a,S=1}$
unmeasured confounding $Y_a \perp\!\!\!\perp A \mid C, U_c$	exposure misclassification $\Pr(Y=0 \mid a, c) \approx 1$ $\Pr(A=0 \mid y, c) \approx 1$	general selection bias $Y \perp\!\!\!\perp S \mid A^*, C, U_s$	$g\left(\mathrm{RR}_{AU_c}, \mathrm{RR}_{U_cY}\right) \times \mathrm{RR}_{YA^*\mid a} \times$ $g\left(\mathrm{RR}_{U_sY\mid A^*=1}, \mathrm{RR}_{SU_s\mid A^*=1}\right) \times$ $g\left(\mathrm{RR}_{U_sY\mid A^*=0}\mathrm{RR}_{SU_s\mid A^*=0}\right)$
unmeasured confounding $Y_a \perp\!\!\!\perp A \mid S=1, C, U_c, U_s$	selected population	exposure misclassification $\Pr(Y=0 \mid a, c, S=1) \approx 1$ $\Pr(A=0 \mid y, c, S=1) \approx 1$	$g\left(\mathrm{RR}_{AU_{sc}\mid S=1}, \mathrm{RR}_{U_{sc}Y\mid S=1}\right) \times \mathrm{RR}_{YA^*\mid a}$

Example

We illustrate the use of the multiple-bias bound to assess possible bias in the study by Omoni and colleagues regarding the effect of HIV status on wasting.[100] Wasting is defined by weight-for-length Z-score of -2 or below and is a rare outcome, so we can interpret the reported OR of 6.75 (95% CI, 2.79, 16.31) as an approximate risk ratio. Since we have no reason to believe that misclassification of wasting was differential by exposure status (i.e., child or mother HIV status), and non-differential outcome misclassification would on average bias toward the null in this situation,[80] we will focus on unmeasured confounding and selection bias in this example.

The choice of whether to participate in the trial, and therefore in the analysis in question, may have been influenced by prior maternal HIV status. For example, people with HIV infection may be hesitant to enroll due to stigma regarding infection, or fear of confirming their status. Other factors may affect enrollment as well: parents with food insecurity may be more likely to enroll in a vitamin-supplementation trial than those without, if they think it will improve their children's nutrition. This benefit could outweigh the hesitancy for some, resulting in selection bias: if a mother in the study is living with HIV, it is likely that her family is also food insecure, making her child more at risk of wasting. Participation in the trial is therefore a collider in a directed acyclic graph describing these relationships, as shown in Figure 3.1A. Similarly, there are factors that are associated with HIV status that may also affect wasting; if these are not on the causal pathway, we may be worried about unmeasured confounding. The authors did not adjust for parity or marital status, though they report that primiparous women were less likely to have HIV, as were married women. We may be concerned that children in single-parent households and those with more siblings are at higher risk of wasting. To demonstrate interpretation of the bound, we will propose values for the parameters describing the strength of these relationships based on the data presented in the original article as well as our background knowledge. Suppose that the most vulnerable in the population were more likely to participate in the trial, and thus that wasting is more likely in children of participants than of non-participants, both among those

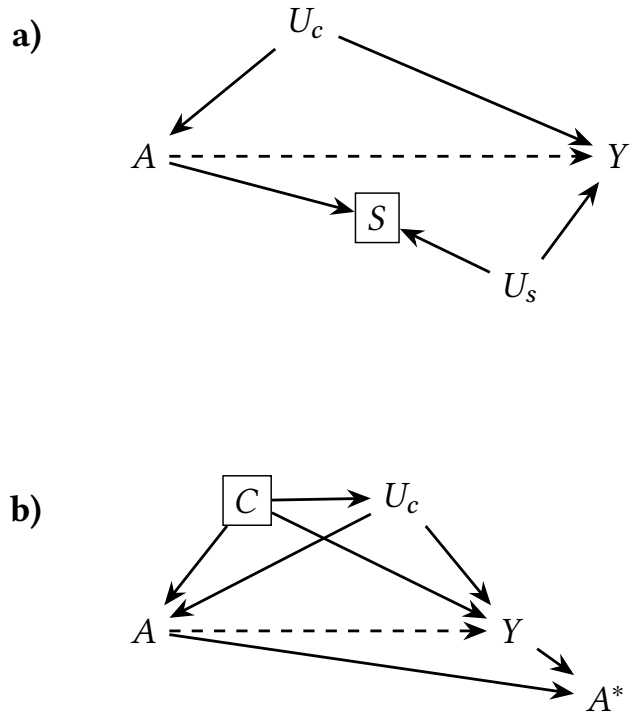

Figure 3.1: *Directed acyclic graphs depicting the examples described in the text.* ***a)*** *This DAG depicts unmeasured confounding (due to U_c) and selection bias (due to U_s). In the graph, the assumptions $Y_a \perp\!\!\!\perp A \mid U_s, S = 1$ and $Y_a \perp\!\!\!\perp A \mid U_c$ are met. This corresponds to the first example in the text, where A indicates HIV infection, U_c family factors including parity and marital status, S participation in the trial, U_s food insecurity, and Y wasting.* ***b)*** *This DAG depicts unmeasured confounding (due to U_c) and differential misclassification of the exposure (due to the $Y \rightarrow A^*$ edge). In the graph, the assumption $Y_a \perp\!\!\!\perp A \mid C, U_c$ is met. This corresponds to the second example in the text, where A indicates vitamin consumption during pregnancy, A^* reported vitamin consumption, U_c breastfeeding, C maternal age, race, and education, and Y child leukemia.*

with HIV as well as those without. The assumption that the outcome is more likely in the selected population of both exposure groups allows us to simplify the selection bias component of the bound, so that the bounding factor only relies on two selection terms, as described by Smith and VanderWeele.[95] Suppose now that children of the most food-insecure mothers are 3 times as likely to have extremely low weight-for-length scores than the least likely group, so that $\text{RR}_{U_s Y|A=1} = 3$, and that the mothers with HIV infection in the study compared to those not in the study are twice as likely to be food insecure, so that $\text{RR}_{SU_s|A=1} = 2$.

Although the odds ratios from this study were not adjusted for parity and marital status, the authors reported proportions of these characteristics stratified by exposure,[100] which can aid in coming up with a reasonable value for RR_{AU_c}. For example, suppose we estimate that 3% of the women whose infants are infected with HIV are multiparous and unmarried, but that this is true of 7% of the women without HIV. If this is the family situation with the largest disparity between exposure groups, then we can specify $\text{RR}_{AU_c} = 2.3$. Now suppose that children in these most precarious families have 2.5 times the risk of wasting than those in the least precarious, so that $\text{RR}_{U_c Y} = 2.5$.

Then we can calculate the bound as $\frac{3\times2}{3+2-1} \times \frac{2.3\times2.5}{2.3+2.5-1} = 2.27$. If those are the only sources of selection bias and unmeasured confounding, and there is no measurement error, then this amount of bias cannot fully explain the approximate observed $\text{RR}_{AY}^{\text{obs}}$ of 6.75, since $6.75/2.27 = 2.97$. Of course, this observed value is subject to statistical uncertainty, so we can also consider the lower limit of the confidence interval, 2.79. If the proposed parameter values hold, then even in the worst case scenario, $\text{RR}_{AY}^{\text{true}}$ is still consistent with $2.79/2.27 = 1.23$, an increase of about 23% in risk of wasting at 2 years of age due to HIV infection. However, the parameter values we used represent only a single reasonable choice. We might proceed by exploring a range of values, as we demonstrate below when we introduce software.

Exposure misclassification

When differential exposure misclassification is a concern, we can derive a similar bound under similar assumptions. However, unlike the bound for outcome misclassification, the bound for exposure misclassification that is employed applies to the odds ratio, not the risk ratio, and the sensitivity parameters are also not themselves risk ratios.[96] We therefore cannot factor the observed risk ratio as in the previous section. However, for a sufficiently rare outcome, odds ratios approximate risk ratios, which allows for some progress.

In this section, $\text{RR}^{\text{obs}}_{A^*Y} = \frac{\Pr(Y=1|A^*=1,S=1,c)}{\Pr(Y=1|A^*=0,S=1,c)}$ refers to the observed (approximate) risk ratio under exposure misclassification, when the outcome is rare in the selected population. Denote with $\text{OR}_{A^*Y|y,S=1}$ the largest out of the false-positive odds ratio $\left\{f_1'/f_0'\right\} / \left\{(1-f_1')/(1-f_0')\right\}$, the sensitivity odds ratio $\left\{s_1'/s_0'\right\} / \left\{(1-s_1')/(1-s_0')\right\}$, the correct classification ratio $\{s_1'/s_0'\}/\{(1-f_1')/(1-f_0')\}$, and incorrect classification ratio $\left\{f_1'/f_0'\right\} / \left\{(1-s_1')/(1-s_0')\right\}$, where $f_y' = \Pr(A^* = 1 \mid Y = y, A = 0, S = 1, c)$ and $s_y' = \Pr(A^* = 1 \mid Y = y, A = 1, S = 1, c)$. Then the following bound holds approximately; i.e., to the extent that the odds ratio approximates the risk ratio.

Result 2:

If $\Pr(Y = 0 \mid A^* = a, S = 1, c) \approx 1$ and $\Pr(Y = 0 \mid A = a, S = 1, c) \approx 1$, then:

$$\text{RR}^{\text{obs}}_{A^*Y}/\text{RR}^{\text{true}}_{AY} \leq \text{BF}_m \times \text{BF}_s \times \text{BF}_c$$

where $\text{BF}_m = \text{OR}_{A^*Y|y,S=1}$ and BF_s and BF_c are as previously defined.

This result is summarized in the fourth row of Table 3.1, as are extensions involving exposure misclassification.

Example

We can jointly assess the magnitude of bias due to differential recall of vitamin use and unmeasured confounding in the study of leukemia risk by Ross and colleagues, in which $\text{RR}^{\text{obs}}_{A^*Y} = 0.51$ (95% CI 0.30, 0.89), by proposing realistic values for the bias parameters. A probabilistic bias analysis for

misclassification was previously done in relation to this study, in which Jurek et al. conducted a literature search for validation studies of multivitamin use during the periconceptional period.[105] They found no pertinent articles, and instead used expert knowledge and bounds from the data (e.g., by assuming correct classification is better than chance) to propose distributions for false negative and false positive probabilities for the cases and controls, which we can use to inform our choice of parameters. Because we think the case-control differential in false negatives is stronger than that for false positives, we might choose that $\Pr(A^* = 0 \mid Y = 1, A = 1) = 0.15$ and $\Pr(A^* = 0 \mid Y = 0, A = 1) = 0.1$ to compute BF'_m. Since we are dealing with a possibly protective factor, however, and the bound is greater than 1 by definition, we reverse the coding of the exposure to reflect that the original estimate of $\text{RR}^{\text{obs}}_{A^*Y} = 0.51$ represents a $1/0.51 = 1.96$-fold increase in risk associated with *not* taking vitamins. Therefore, $f'_1 = 0.15$ and $f'_0 = 0.10$, and $\text{BF}'_m = 1.59$.

Jurek et al.'s probabilistic bias analysis used the crude 2-by-2 table from the original article, so did not take into account even the few measured confounders.[105] However, even those measured confounders would likely not be sufficient to control for confounding by healthy lifestyle, as there is evidence that other healthy behaviors are associated with leukemia. For example, a recent meta-analysis found that not breastfeeding compared to breastfeeding for at least 6 months was associated with an increase in acute lymphoblastic leukemia risk by a factor of 1.22.[106] Using breastfeeding as a proxy for healthy lifestyle, for the unmeasured confounding parameters, we will take $\text{RR}_{U_cY} = 1.22$ and $\text{RR}_{AU_c} = 2$, suggesting that children who weren't breastfed are 1.22 times as likely to get leukemia, and that mothers who take multivitamins are twice as likely to breastfeed than those who do not. A directed acyclic graph depicting this example is shown in Figure 3.1B.

Using these values, we find that $1.59 \times \frac{1.22 \times 2}{1.22 + 2 - 1} = 1.75$, indicating that the observed risk ratio may be biased by a factor of 1.75 if the differential misclassification and unmeasured confounding were of the strengths we proposed. Since we are dealing with a possibly protective factor, we multiply the observed estimate of 0.51 and its confidence interval (95% CI 0.30, 0.89) by the bound (or equivalently divide the reverse-coded estimate of 1.96 by the bound), resulting in a bias-adjusted

estimate and confidence interval of 0.89 (95% CI 0.52, 1.56). Unlike the Jurek et al. sensitivity analysis,[105] which found that results were largely unchanged by exposure misclassification, we have focused specifically on a situation in which misclassification is differential by outcome, and have additionally taken both measured and unmeasured confounding into account. Doing so indicates that the results may be sensitive to misclassification and uncontrolled confounding, as can be seen if the chosen parameter values are thought to be reasonable.

Inference in the selected population

Results 1 and 2 are derived with respect to the true causal effect in the total population, despite possible selection bias. In other situations, we may only be interested in the existence and magnitude of a causal effect in the selected population. In this case, our estimand of interest is $\text{RR}^{\text{true}}_{AY|S=1} = \frac{\Pr(Y_1|S=1,c)}{\Pr(Y_0|S=1,c)}$. If only selection bias is present, one can derive a bound under the assumption that $Y_a \perp\!\!\!\perp A \mid S = 1, c, U_s$.[95] In the present context, we additionally accommodate unmeasured confounding and measurement error. Consider unmeasured confounding by U_c such that it is only the case that $Y_a \perp\!\!\!\perp A \mid S = 1, c, U_s, U_c$. Therefore, we must consider the vector of factors causing selection bias and unmeasured confounding $U_{sc} = (U_s, U_c)$. Define the sensitivity parameters $\text{RR}_{U_{sc}Y} = \max_a \frac{\max_u \Pr(Y=1|A=a,c,U_{sc}=u)}{\min_u \Pr(Y=1|A=a,c,U_{sc}=u)}$ and $\text{RR}_{AU_{sc}} = \max_u \frac{\Pr(U_{sc}=u|A=1,c)}{\Pr(U_{sc}=u|A=0,c)}$. Then under outcome misclassification, we have the following bound.

Result 3:

If $Y_a \perp\!\!\!\perp A \mid S = 1, c, U_c, U_s$, then:

$$\text{RR}^{\text{obs}}_{AY^*} / \text{RR}^{\text{true}}_{AY|S=1} \leq \text{BF}_m \times \text{BF}_{sc}$$

where BF_m is defined as in Result 1, and $\text{BF}_{sc} = g\left(\text{RR}_{U_{sc}Y}, \text{RR}_{AU_{sc}}\right)$. These latter parameters now refer to the maximum risk ratio for the outcome among the selected comparing any two levels of any of U_s and U_c, and the maximum ratio for any joint level of U_s and U_c comparing exposed to

unexposed, among the selected. This bound holds under exposure misclassification with a rare outcome in the selected population as well, with $\text{BF}'_m = \text{OR}_{A^*Y|y,S=1}$.

This result is summarized in the third row of Table 3.1.

3.4 Software

The R package EValue[107] allows for easy calculation of the multiple-bias bounds for various combinations of biases and assumptions, including all those presented in Table 3.1, as well as the possible simplifications to the selection bias bound, as in the first example. The function `multi_bias()` creates a set of biases according to the user's specifications. The user can then input this object along with a proposed set of parameter values to the `multi_bound()` function to calculate a bound.

For example, the biases in the HIV example can be set with `HIV_biases <- multi_bias(confounding(), selection("general", "increased risk"))`. The command to calculate the bound is then `multi_bound(biases = HIV_biases, RRAUc = 2.3, RRUcY = 2.5, RRUsYA1 = 3, RRSUsA1 = 2)`. Similarly, for the vitamins-leukemia example, the biases are set with `leuk_biases <- multi_bias(confounding(), misclassification("exposure", rare_outcome = TRUE, rare_exposure = FALSE))` and the bound command is `multi_bound(biases = leuk_biases, RRAUc = 2, RRUcY = 1.22, ORYAa = 1.59)`.

These functions can be used to prepare a table or figure of bounded bias-adjusted estimates across a range of proposed parameter values. For example, Table 3.2 shows the upper bound for the estimate of the protective effect of multivitamin use on leukemia across various values for RR_{U_cY} and RR_{AU_c} (in the columns and rows), and for two values of the misclassification ratio $\text{OR}_{A^*Y|y}$ above and below the diagonal. We can use this table to describe multiple scenarios under which it would be possible for the true effect to be null. We can also see that even if, for example, the prevalence of the unmeasured confounder, or set of confounders, differs greatly between

Table 3.2: *Corrected estimates for the effect of multivitamin use in pregnancy on childhood leukemia, taking into account unmeasured confounding and recall bias. The original estimate was 0.51. Corrected estimates are arranged in rows and columns by the parameters defining the unmeasured confounding, RR_{AU_c} and RR_{U_cY}. The two parameters are interchangeable with respect to the bound, so a table of estimates corrected only for unmeasured confounding would be symmetric. However, the estimates in the upper and lower triangles have been corrected by misclassification parameters of different magnitudes. Below the diagonal, the misclassification ratio is assumed to be 1.25; above the diagonal, it is assumed to be 1.5.*

RR_{AU_c} \ RR_{U_cY}	1.25	1.5	1.75	2	2.25	2.5	2.75	3
1.25	0.80 0.66	0.82	0.84	0.85	0.86	0.87	0.88	0.88
1.5	0.68	0.86 0.72	0.89	0.92	0.94	0.96	0.97	0.98
1.75	0.70	0.74	0.94 0.78	0.97	1.00	1.03	1.05	1.07
2	0.71	0.76	0.81	1.02 0.85	1.06	1.09	1.12	1.15
2.25	0.72	0.78	0.84	0.88	1.11 0.92	1.15	1.18	1.22
2.5	0.72	0.80	0.86	0.91	0.96	1.20 1.00	1.24	1.27
2.75	0.73	0.81	0.88	0.94	0.99	1.03	1.29 1.07	1.33
3	0.74	0.82	0.89	0.96	1.01	1.06	1.11	1.38 1.15

consumers and non-consumers of multivitamins (e.g., $RR_{AU_c} = 3$), a relatively small association between the unmeasured confounder and leukemia (e.g., $RR_{U_cY} = 1.25$) and between vitamin use and misclassifcation (e.g., $OR_{A^*Y|y} = 1.25$) would at most lead to a bias-adjusted estimate of 0.74. More examples are available in the eAppendix, and the package documentation is available online.

3.5 Discussion

We have described an approach to sensitivity analysis that we hope can help bridge the gap between complex methods that require specifying many parameters and making restrictive

assumptions, and simpler methods that allow for assessment of only one type of bias at a time. The multiple-bias bound can be used to simultaneously consider the possible effects of biases that are of different strengths. Researchers can propose values for the parameters based on background knowledge, validation studies, or simply hypothetical situations, and assess the minimum possible true risk ratio that would be compatible if the observed value were affected by biases of that magnitude. When planning for future research, the bound can be used to compare the effects of biases within a given situation and prioritize more extensive confounder assessment, a more valid sampling/inclusion scheme, and/or better measurement techniques if resource constraints or data collection options force one to choose among them. It may also show that certain improvements to study design are futile; if the amount of an unavoidable bias greatly attenuates the anticipated risk ratio estimate, investing resources into reducing another type of bias may not be worth it.[108,109]

There are a number of caveats and limitations to this approach. Although the calculations involved in our approach are simple, the entire process of assessing bias should not be. Importantly, it should be specific to the study design, the available data, and the research question; values for the sensitivity parameters are meaningless without a frame of reference. Indeed, critiques of the E-value for unmeasured confounding have emphasized the importance of clearly specifying the confounder, or set of confounders, that have not been measured.[110–112] The same should be true for factors potentially causing selection bias, or the reason behind possible differential misclassification. Unmeasured confounders could be anything from a single missing risk factor to the "ultimate covariate,"[113] the variable encoding an individual's causal type. Misclassification may be negligible or close to non-differential, or as bad as chance in one or another group; it is up to researchers and readers to assess the plausibility of these situations with respect to a given study and what was conditioned on in the analysis, and then assess how much bias they would create. Like any tool, the multiple-bias bound can be misused; we encourage researchers to not be careless, due to its apparent simplicity, but rather to be thoughtful in its use.

Additionally, in avoiding certain assumptions, we have necessarily invoked others. In particular, the bounds we propose describe a "worst-case scenario" for the bias; in almost all settings, the

actual bias will be smaller than the bound. For example, for the actual bias to obtain the bound assumes that the unmeasured confounder has the distribution that maximizes confounding, given the two parameters defining it.[114] The same is true of selection bias and misclassification, e.g., the general selection bias bound implies that outcomes and exposures in the non-selected group are distributed to result in the most possible bias. This is of course necessary for a bound to be a bound, but many realistic conditions would not result in as much bias, and the bound should be interpreted as the bias that *could* result from parameters of a given magnitude, not that necessarily *would* result. The few assumptions that are required for the bound to hold may not be reasonable in all settings; for example, the general selection bias assumption is unlikely to hold in case-control studies. In addition, the interpretation of the bias-adjusted confidence interval pertains to the application of the adjustment to repeated samples with the same sets of biases; if the biases were truly resolved in the design or analysis, the bounds would differ.

Finally, while we have suggested two possible orderings for factoring the bias, others that take into account, for example, misclassification that is also differential by an unmeasured confounder, are possible. We have presented results for risk ratios, which can in many cases be extended to odds ratios. However, our bound for exposure misclassification relies on a rare outcome assumption that limits its use and results in an approximate bound. Because we are never sure of the true values of the parameters that make up the bound, and the bound represents a worst-case scenario not likely to hold anyway, this approximation is not likely to meaningfully affect interpretation. Further work could be done to extend this approach to risk differences or mean differences, which may be especially challenging because the bounds are more frequently non-informative.[115] Other approaches exist to quantify as simply as possible unmeasured confounding in linear or probit models,[116–119] but to our knowledge they have not yet been extended to multiple biases.

There is no single solution to the problem of bias in epidemiologic research. Some biases can be corrected at the design phase, others in the main analysis, but the assessment of what bias may remain should be a regular component of any study that attempts to quantify causal effects.

The multiple-bias bound can make it simpler to do so, and we hope to encourage thoughtful consideration of multiple sources of bias in epidemiologic research.

A

Appendix to Chapter 1

A.1 Additional tables and figures

Table A.1: *Comparison of descriptive characteristics (n (%)) of IRCEP participants who provided outcome data and those who did not, but whose pregnancies were presumed completed (i.e., lost to follow-up).*

	No outcome data N = 5,225	Provided outcome data N = 5,848	Total N = 11,073
Enrollment			
Prospective	5,023 (96%)	392 (6.7%)	5,415 (49%)
Retrospective	202 (3.9%)	5,456 (93%)	5,658 (51%)
Prospective			
COVID-19 positive	2,662 (53%)	155 (40%)	2,817 (52%)
Gestational age at enrollment[a]	31 (25, 36)	34 (31, 37)	32 (26, 36)
Gestational age at enrollment[a]	46 (41, 53)	49 (44, 54)	49 (44, 54)
Retrospective			
COVID-19 positive	56 (28%)	1,018 (19%)	1,074 (19%)
Gestational age at symptom onset/test[a]	25 (18, 32)	27 (20, 33)	25 (18, 32)
Gestational age at symptom onset/test[a]	37.1 (28.3, 39.0)	38.1 (35.6, 39.3)	38.0 (35.4, 39.3)
COVID-19 severity			
Asymptomatic	263 (9.7%)	198 (17%)	461 (12%)
Mild	985 (36%)	383 (33%)	1,368 (35%)
Moderate	1,345 (49%)	463 (39%)	1,808 (46%)

Table A.1: *Comparison of descriptive characteristics (n (%)) of IRCEP participants who provided outcome data and those who did not, but whose pregnancies were presumed completed (i.e., lost to follow-up). (continued)*

	No outcome data N = 5,225	Provided outcome data N = 5,848	Total N = 11,073
Severe	125 (4.6%)	129 (11%)	254 (6.5%)
COVID-19 diagnosis/test type			
Negative	2,507 (48%)	4,675 (80%)	7,182 (65%)
Positive by antibodies only	268 (5.1%)	150 (2.6%)	418 (3.8%)
Positive by throat/nose swab	2,158 (41%)	863 (15%)	3,021 (27%)
Positive clinically only	292 (5.6%)	160 (2.7%)	452 (4.1%)
Reason for COVID-19 test			
Symptoms	2,547 (49%)	1,011 (17%)	3,558 (32%)
Contact tracing/risk zone travel	1,220 (23%)	635 (11%)	1,855 (17%)
Surveillance (healthy)	653 (13%)	1,696 (29%)	2,349 (21%)
Other/none	803 (15%)	2,506 (43%)	3,309 (30%)
Age[a]	30.0 (27.0, 34.0)	31.0 (27.0, 34.0)	31.0 (27.0, 34.0)
Healthcare coverage	3,364 (84%)	5,038 (89%)	8,402 (87%)
Pre-existing condition	503 (14%)	779 (14%)	1,282 (14%)
Primiparous	1,612 (43%)	2,589 (46%)	4,201 (45%)
Pre-pregnancy BMI			
<25	1,593 (48%)	2,445 (46%)	4,038 (47%)
25-30	884 (27%)	1,412 (26%)	2,296 (27%)
≤ 30	811 (25%)	1,502 (28%)	2,313 (27%)
Continent			
Africa	227 (4.3%)	290 (5.0%)	517 (4.7%)
Asia	446 (8.5%)	301 (5.1%)	747 (6.7%)
Europe	1,364 (26%)	2,012 (34%)	3,376 (30%)
North America	1,499 (29%)	2,419 (41%)	3,918 (35%)
South America	1,689 (32%)	825 (14%)	2,514 (23%)

[a] Median (interquartile range)

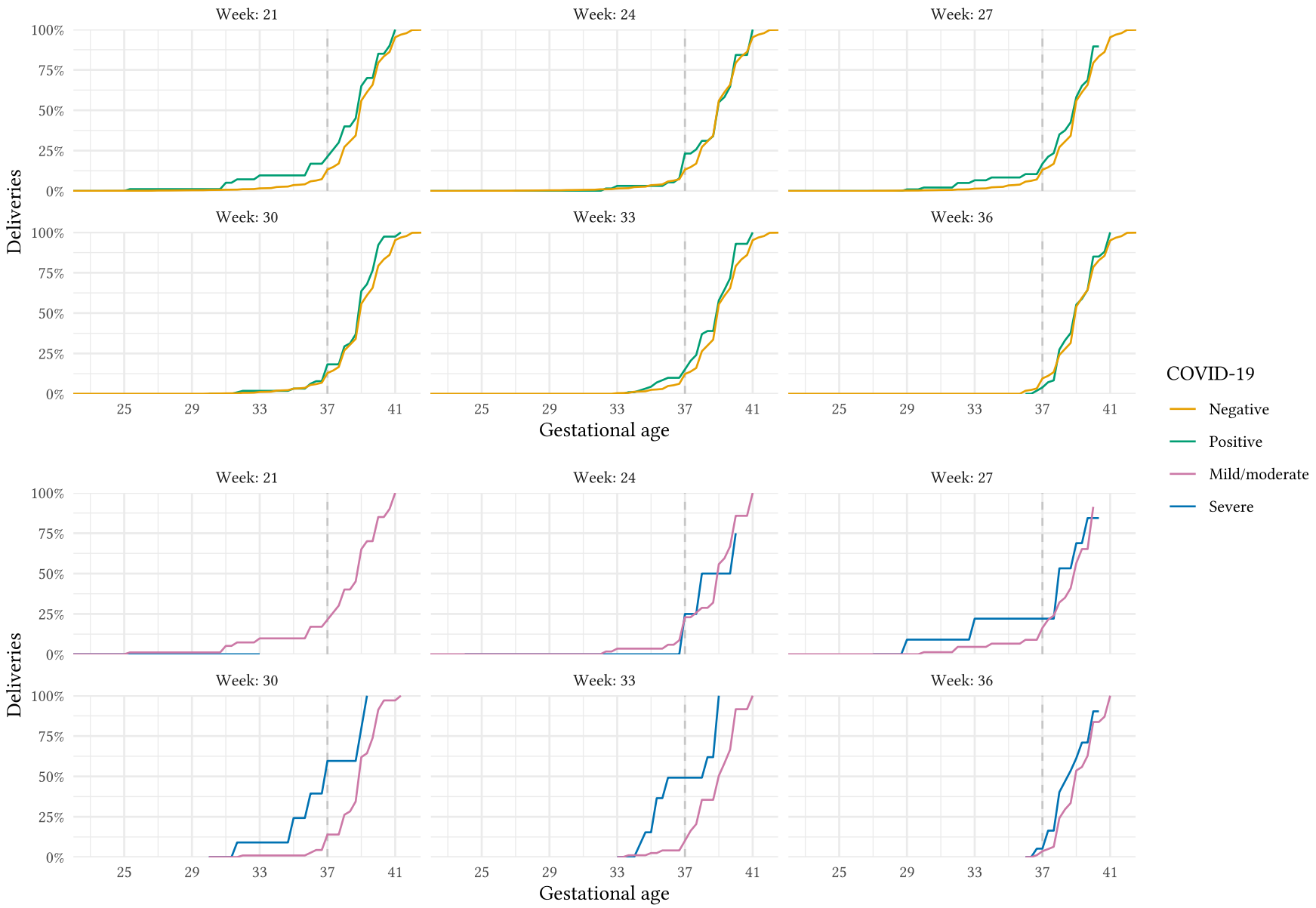

Figure A.1: *Unadjusted cumulative probabilities of delivery after COVID-19 in selected weeks of gestation. COVID-19 negative individuals in a given week are those who are still pregnant at that week. Week 20 refers to all infections at or before week 20.*

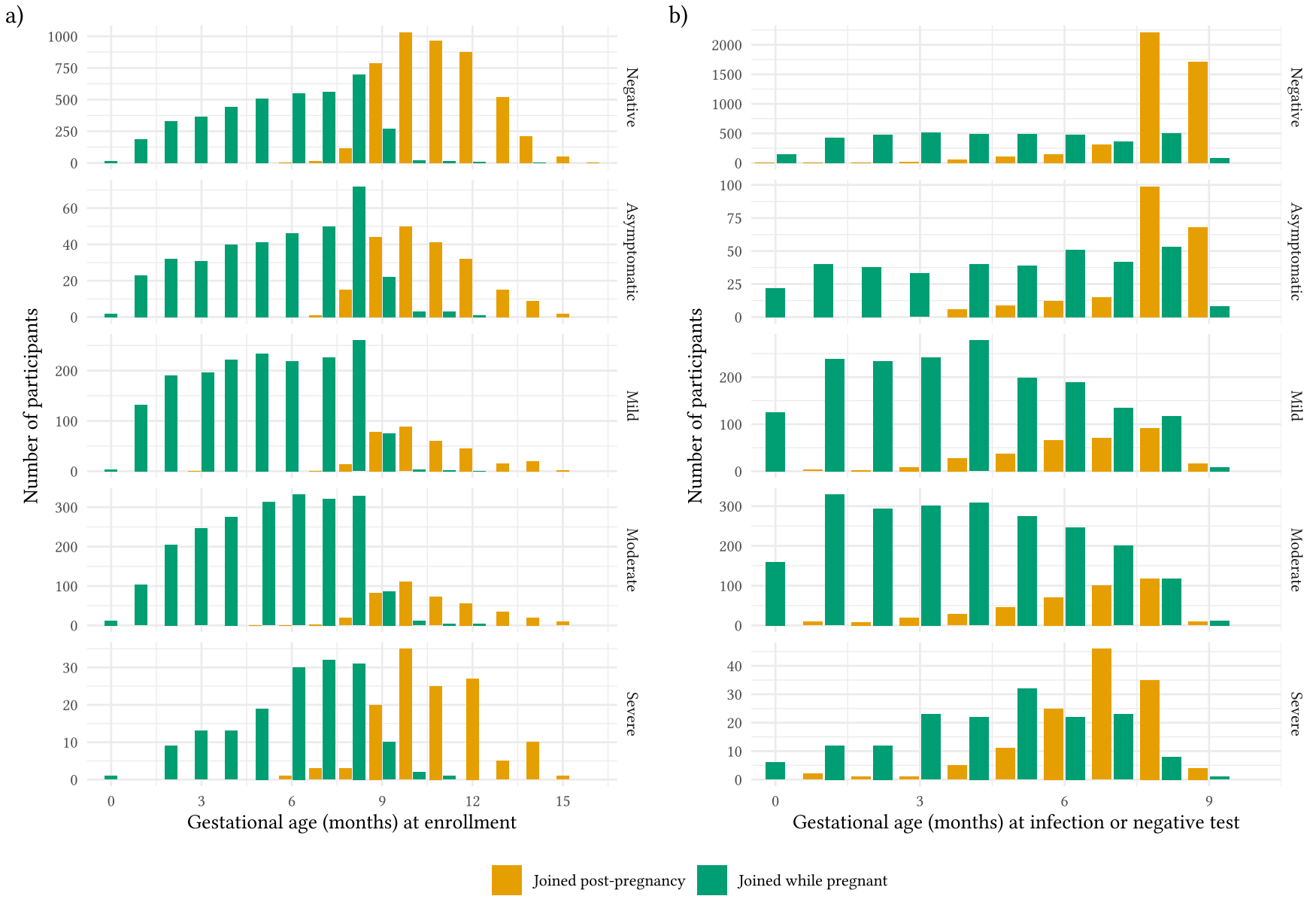

Figure A.2: *Gestational age at a) enrollment and b) time of symptom onset or test (if negative or positive, symptomatic), for participants who joined while pregnant and within 6 months after pregnancy, stratified by COVID-19 severity.*

Table A.2: *Estimates of standardized risks of spontaneous preterm delivery and risk ratios comparing COVID-19 positive vs. negative and severe vs. mild/moderate.*

	Standardized risks		Risk ratio	Standardized risks		Risk ratio
	Negative	Positive	Positive vs. negative	Mild/moderate	Severe	Severe vs. mild/moderate
Week 20	6.6% (5.8, 7.4)	6.0% (4.4, 7.9)	0.9 (0.7, 1.2)	6.1% (4.5, 8.1)	4.7% (2.3, 8.0)	0.8 (0.4, 1.3)
Week 21	6.6% (5.8, 7.4)	6.0% (4.5, 7.7)	0.9 (0.7, 1.2)	6.1% (4.5, 7.8)	5.1% (2.6, 8.4)	0.8 (0.5, 1.3)
Week 22	6.6% (5.8, 7.4)	6.1% (4.4, 7.8)	0.9 (0.7, 1.2)	6.2% (4.5, 7.9)	5.5% (3.0, 8.9)	0.9 (0.5, 1.4)
Week 23	6.6% (5.8, 7.3)	6.3% (4.6, 8.0)	1.0 (0.7, 1.3)	6.3% (4.5, 8.0)	6.1% (3.2, 9.7)	1.0 (0.5, 1.5)
Week 24	6.5% (5.8, 7.3)	6.4% (4.7, 8.2)	1.0 (0.7, 1.3)	6.4% (4.7, 8.2)	6.8% (3.5, 10.5)	1.1 (0.6, 1.6)
Week 25	6.5% (5.8, 7.3)	6.6% (5.0, 8.4)	1.0 (0.8, 1.3)	6.6% (5.0, 8.4)	7.5% (4.1, 11.3)	1.1 (0.6, 1.7)
Week 26	6.5% (5.7, 7.2)	6.8% (5.2, 8.6)	1.1 (0.8, 1.4)	6.7% (5.1, 8.5)	8.3% (4.7, 12.2)	1.2 (0.7, 1.9)
Week 27	6.4% (5.7, 7.2)	7.0% (5.4, 8.8)	1.1 (0.8, 1.4)	6.9% (5.2, 8.6)	9.1% (5.2, 13.3)	1.3 (0.7, 2.0)
Week 28	6.4% (5.6, 7.1)	7.1% (5.4, 8.9)	1.1 (0.9, 1.4)	6.9% (5.2, 8.7)	9.9% (5.6, 14.3)	1.4 (0.8, 2.1)
Week 29	6.3% (5.6, 7.1)	7.1% (5.4, 8.8)	1.1 (0.9, 1.4)	6.9% (5.2, 8.6)	10.6% (5.8, 15.4)	1.5 (0.9, 2.3)
Week 30	6.3% (5.5, 7.0)	7.0% (5.3, 8.8)	1.1 (0.8, 1.4)	6.8% (5.0, 8.5)	11.2% (6.3, 16.7)	1.7 (0.9, 2.6)
Week 31	6.1% (5.4, 6.8)	6.8% (5.1, 8.5)	1.1 (0.8, 1.4)	6.5% (4.8, 8.3)	11.5% (6.6, 17.4)	1.8 (1.0, 2.7)
Week 32	6.0% (5.2, 6.6)	6.4% (4.8, 8.2)	1.1 (0.8, 1.4)	6.1% (4.4, 7.9)	11.7% (6.4, 17.9)	1.9 (1.1, 2.9)
Week 33	5.7% (5.0, 6.4)	5.9% (4.3, 7.7)	1.0 (0.7, 1.4)	5.6% (4.0, 7.2)	11.5% (6.2, 18.0)	2.1 (1.1, 3.2)
Week 34	5.3% (4.7, 5.9)	5.3% (3.8, 6.8)	1.0 (0.7, 1.3)	4.9% (3.5, 6.4)	10.9% (5.6, 17.6)	2.2 (1.2, 3.5)
Week 35	4.6% (4.0, 5.2)	4.3% (3.0, 5.6)	0.9 (0.6, 1.2)	4.0% (2.7, 5.2)	9.5% (4.8, 16.0)	2.4 (1.3, 3.9)
Week 36	3.2% (2.8, 3.6)	2.8% (1.9, 3.7)	0.9 (0.6, 1.2)	2.5% (1.7, 3.4)	6.6% (3.2, 11.7)	2.6 (1.3, 4.4)

Table A.3: *Estimates of standardized risks of induced preterm delivery and risk ratios comparing COVID-19 positive vs. negative and severe vs. mild/moderate.*

	Standardized risks		Risk ratio	Standardized risks		Risk ratio
	Negative	Positive	Positive vs. negative	Mild/moderate	Severe	Severe vs. mild/moderate
Week 20	3.2% (2.5, 3.9)	3.9% (2.3, 5.4)	1.2 (0.8, 1.8)	3.8% (2.3, 5.4)	4.4% (2.1, 8.0)	1.1 (0.6, 2.0)
Week 21	3.2% (2.5, 3.9)	3.9% (2.4, 5.4)	1.2 (0.8, 1.8)	3.8% (2.3, 5.4)	4.8% (2.3, 8.4)	1.2 (0.7, 2.2)
Week 22	3.2% (2.4, 3.9)	4.0% (2.5, 5.5)	1.3 (0.8, 1.8)	3.9% (2.4, 5.4)	5.2% (2.7, 8.8)	1.3 (0.8, 2.3)
Week 23	3.1% (2.4, 3.9)	4.1% (2.6, 5.6)	1.3 (0.8, 1.9)	4.0% (2.5, 5.5)	5.8% (3.1, 9.5)	1.5 (0.8, 2.5)
Week 24	3.1% (2.4, 3.9)	4.3% (2.8, 5.7)	1.4 (0.9, 1.9)	4.1% (2.6, 5.6)	6.5% (3.6, 10.5)	1.6 (0.9, 2.6)
Week 25	3.1% (2.4, 3.8)	4.4% (3.0, 5.8)	1.4 (0.9, 2.0)	4.3% (2.7, 5.7)	7.2% (4.1, 11.4)	1.7 (1.0, 2.8)
Week 26	3.1% (2.4, 3.8)	4.6% (3.1, 6.0)	1.5 (1.0, 2.0)	4.4% (2.8, 5.8)	8.0% (4.6, 12.4)	1.8 (1.1, 3.2)
Week 27	3.1% (2.4, 3.8)	4.8% (3.2, 6.2)	1.5 (1.1, 2.1)	4.5% (3.0, 5.9)	8.9% (5.2, 13.9)	2.0 (1.2, 3.5)
Week 28	3.1% (2.4, 3.8)	4.9% (3.2, 6.4)	1.6 (1.1, 2.2)	4.6% (2.9, 6.0)	9.7% (5.7, 15.1)	2.1 (1.3, 3.8)
Week 29	3.0% (2.4, 3.7)	4.9% (3.1, 6.5)	1.6 (1.1, 2.2)	4.6% (2.8, 6.1)	10.5% (6.2, 15.9)	2.3 (1.4, 4.0)
Week 30	3.0% (2.3, 3.7)	4.9% (3.0, 6.5)	1.6 (1.1, 2.2)	4.5% (2.7, 6.1)	11.2% (6.7, 16.9)	2.5 (1.5, 4.3)
Week 31	3.0% (2.3, 3.6)	4.8% (3.0, 6.5)	1.6 (1.0, 2.2)	4.4% (2.5, 6.0)	11.6% (7.0, 17.5)	2.6 (1.6, 4.6)
Week 32	2.9% (2.3, 3.5)	4.6% (2.9, 6.3)	1.6 (1.0, 2.2)	4.2% (2.3, 5.8)	11.9% (7.0, 17.9)	2.8 (1.7, 5.0)
Week 33	2.8% (2.2, 3.4)	4.3% (2.6, 6.0)	1.6 (1.0, 2.2)	3.9% (2.1, 5.4)	11.8% (6.8, 18.1)	3.1 (1.8, 5.6)
Week 34	2.6% (2.0, 3.2)	3.9% (2.3, 5.4)	1.5 (0.9, 2.2)	3.4% (1.9, 4.8)	11.2% (6.3, 17.5)	3.3 (1.9, 6.1)
Week 35	2.2% (1.7, 2.7)	3.2% (1.8, 4.5)	1.4 (0.8, 2.1)	2.8% (1.5, 4.0)	9.9% (5.4, 15.9)	3.6 (2.0, 6.8)
Week 36	1.5% (1.2, 1.9)	2.1% (1.2, 3.1)	1.3 (0.7, 2.0)	1.8% (0.9, 2.7)	6.9% (3.6, 11.4)	3.9 (2.2, 7.6)

Table A.4: *Risk differences for overall, spontaneous, and induced preterm delivery, comparing COVID-19 positive vs. negative and severe vs. mild/moderate.*

	Positive vs. negative	Severe vs. mild/moderate	Positive vs. negative (Spontaneous)	Severe vs. mild/moderate (Spontaneous)	Positive vs. negative (Induced)	Severe vs. mild/moderate (Induced)
Week 20	0.1% (-1.8, 2.4)	-0.8% (-3.3, 3.0)	-0.6% (-2.4, 1.4)	-1.4% (-3.8, 1.6)	0.7% (-0.8, 2.3)	0.6% (-1.5, 3.6)
Week 21	0.2% (-1.9, 2.1)	-0.1% (-2.5, 3.6)	-0.6% (-2.2, 1.2)	-1.0% (-3.5, 2.0)	0.7% (-0.7, 2.2)	0.9% (-1.2, 4.1)
Week 22	0.4% (-1.9, 2.4)	0.7% (-1.7, 4.3)	-0.5% (-2.1, 1.4)	-0.6% (-3.2, 2.6)	0.8% (-0.6, 2.4)	1.3% (-1.0, 4.8)
Week 23	0.7% (-1.7, 3.0)	1.6% (-0.7, 5.1)	-0.3% (-1.9, 1.6)	-0.2% (-2.8, 3.0)	1.0% (-0.5, 2.5)	1.8% (-0.7, 5.2)
Week 24	1.0% (-1.3, 3.2)	2.7% (0.1, 6.3)	-0.1% (-1.7, 1.8)	0.3% (-2.6, 3.7)	1.1% (-0.4, 2.6)	2.3% (-0.3, 6.1)
Week 25	1.4% (-0.9, 3.6)	3.9% (1.1, 7.8)	0.1% (-1.5, 2.0)	0.9% (-2.5, 4.5)	1.3% (-0.2, 2.8)	3.0% (0.1, 6.9)
Week 26	1.9% (-0.3, 3.8)	5.2% (2.1, 9.5)	0.3% (-1.2, 2.2)	1.5% (-2.3, 5.3)	1.5% (0.0, 2.9)	3.6% (0.5, 8.0)
Week 27	2.2% (0.1, 4.0)	6.6% (3.3, 11.4)	0.6% (-1.0, 2.4)	2.2% (-2.0, 6.3)	1.7% (0.2, 3.1)	4.4% (0.9, 9.2)
Week 28	2.5% (0.6, 4.3)	8.1% (4.3, 13.2)	0.7% (-0.9, 2.5)	2.9% (-1.6, 7.3)	1.8% (0.2, 3.3)	5.1% (1.5, 10.6)
Week 29	2.7% (0.7, 4.5)	9.6% (5.3, 15.2)	0.8% (-0.9, 2.6)	3.7% (-1.0, 8.5)	1.9% (0.2, 3.4)	5.9% (2.0, 11.5)
Week 30	2.7% (0.8, 4.6)	11.0% (6.1, 17.0)	0.8% (-1.0, 2.7)	4.4% (-0.4, 9.5)	1.9% (0.2, 3.5)	6.6% (2.4, 12.4)
Week 31	2.5% (0.7, 4.4)	12.3% (6.8, 19.0)	0.6% (-1.1, 2.6)	5.0% (0.1, 10.6)	1.9% (0.1, 3.5)	7.2% (2.9, 13.1)
Week 32	2.2% (0.4, 3.9)	13.3% (7.4, 20.7)	0.5% (-1.2, 2.4)	5.6% (0.4, 11.4)	1.7% (0.1, 3.4)	7.7% (3.3, 13.8)
Week 33	1.8% (0.2, 3.3)	13.9% (7.8, 21.9)	0.2% (-1.5, 2.0)	5.9% (0.8, 11.9)	1.5% (-0.1, 3.1)	8.0% (3.5, 14.1)
Week 34	1.2% (0.0, 2.5)	13.8% (7.7, 22.1)	-0.1% (-1.7, 1.6)	6.0% (0.9, 12.1)	1.3% (-0.2, 2.8)	7.8% (3.4, 13.9)
Week 35	0.6% (-0.3, 1.7)	12.7% (6.9, 20.8)	-0.3% (-1.7, 1.1)	5.6% (1.0, 11.5)	1.0% (-0.4, 2.3)	7.1% (3.1, 12.6)
Week 36	0.1% (-0.5, 0.9)	9.2% (4.9, 15.6)	-0.4% (-1.4, 0.6)	4.1% (0.9, 8.6)	0.5% (-0.4, 1.4)	5.1% (2.2, 9.4)

Table A.5: *Comparison of risk ratios from complete-case (as in the main text) and multiply-imputed analyses.*

Analysis	Positive vs. negative	Moderate vs. mild	Severe vs. mild
Complete-case	1.3 (1.0, 1.7)	1.1 (0.7, 1.7)	2.5 (1.6, 3.9)
Multiply-imputed	1.3 (1.0, 1.6)	1.2 (0.8, 1.8)	2.7 (1.8, 4.2)

A.2 Sensitivity analyses

Table A.6: *Estimates from various sensitivity analyses of the log-linear analysis.*

Analysis	Positive vs. negative	Moderate vs. mild	Severe vs. mild
Original[a]	1.3 (1.0, 1.7)	1.1 (0.7, 1.7)	2.5 (1.6, 3.9)
Exclude clinical[b]	1.4 (1.0, 1.8)	1.0 (0.6, 1.6)	2.4 (1.5, 3.7)
North America[c]	0.8 (0.5, 1.2)	1.5 (0.7, 3.2)	2.9 (1.1, 7.2)
Exclude negative w/in 2 weeks[d]	1.1 (0.7, 1.7)		
Shorter cutoff[e]	1.3 (1.0, 1.7)	1.1 (0.7, 1.7)	2.5 (1.6, 3.9)
Longer cutoff[f]	1.3 (1.0, 1.7)	1.1 (0.7, 1.6)	2.6 (1.7, 4.0)

[a] Analysis as in the main text
[b] Exclude participants with a clinical diagnosis of COVID-19 by no positive test
[c] Include only participants from North America
[d] Exclude participants who tested negative within two weeks prior to delivery
[e] Restrict sample to those with a last menstrual period within 42 (rather than 45) weeks prior
[f] Restrict sample to those with a last menstrual period within 48 (rather than 45) weeks prior

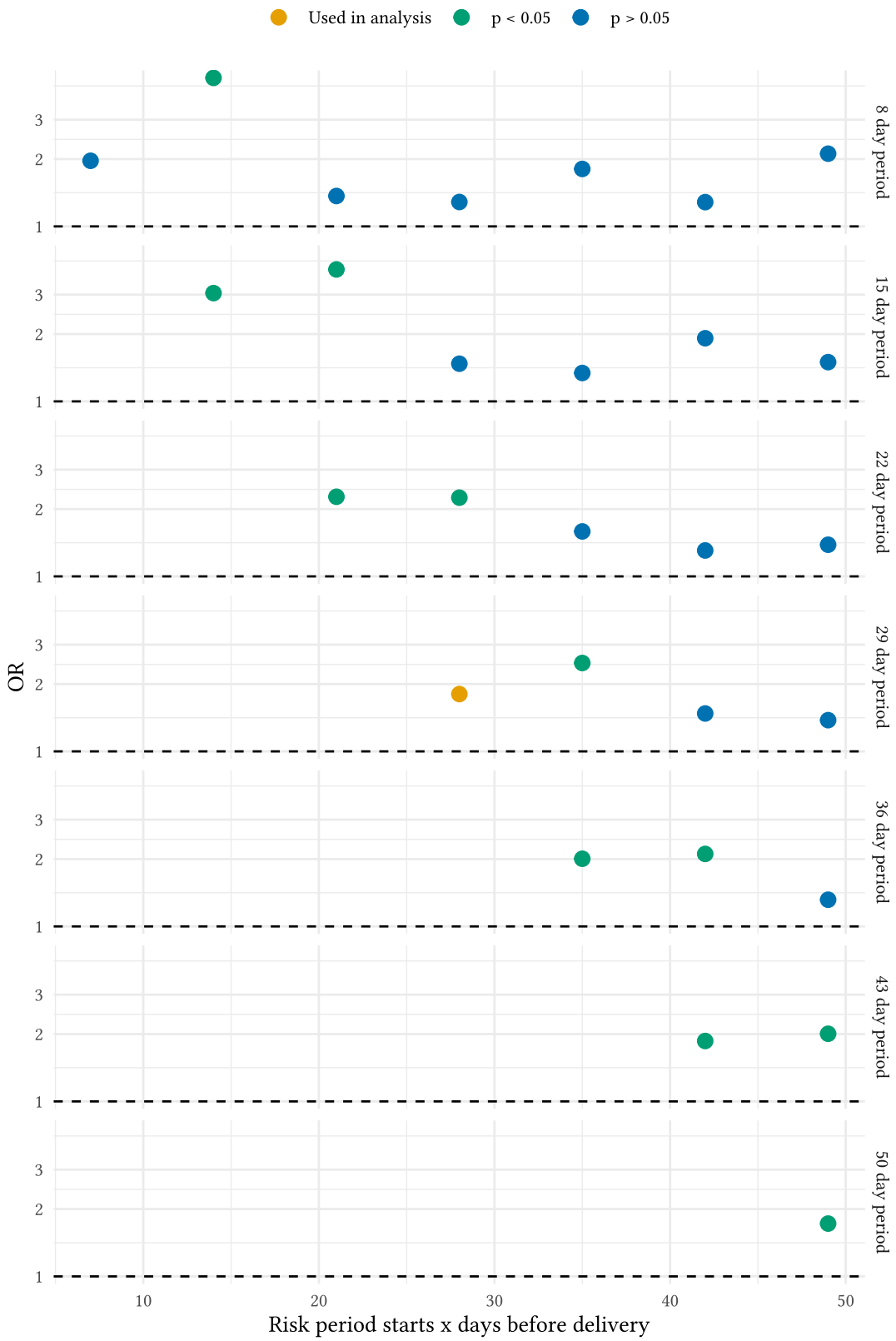

Figure A.3: *Results from varying the risk window (and corresponding size of reference window) in the case-time-control analysis*

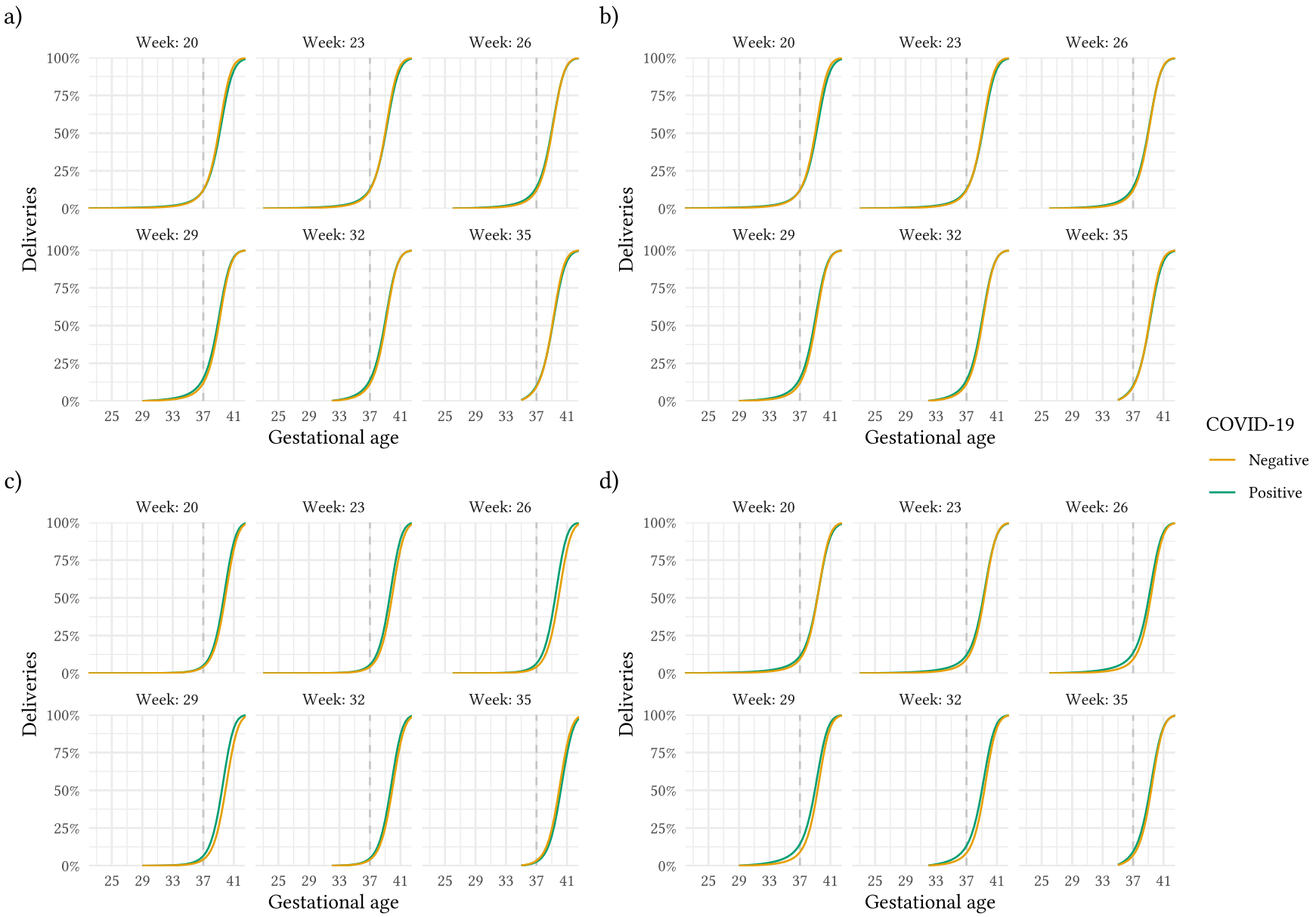

Figure A.4: *Cumulative deliveries across gestation, estimated in various sensitivity analyses (COVID-positive vs. negative): a) Estimating risks separately for mild and moderate severity groups; b) Finer modeling of the discrete-time hazard of delivery; c) Excluding participants who joined after completion of pregnancy; d) Excluding participants who tested negative within two weeks before delivery (i.e., routine delivery testing).*

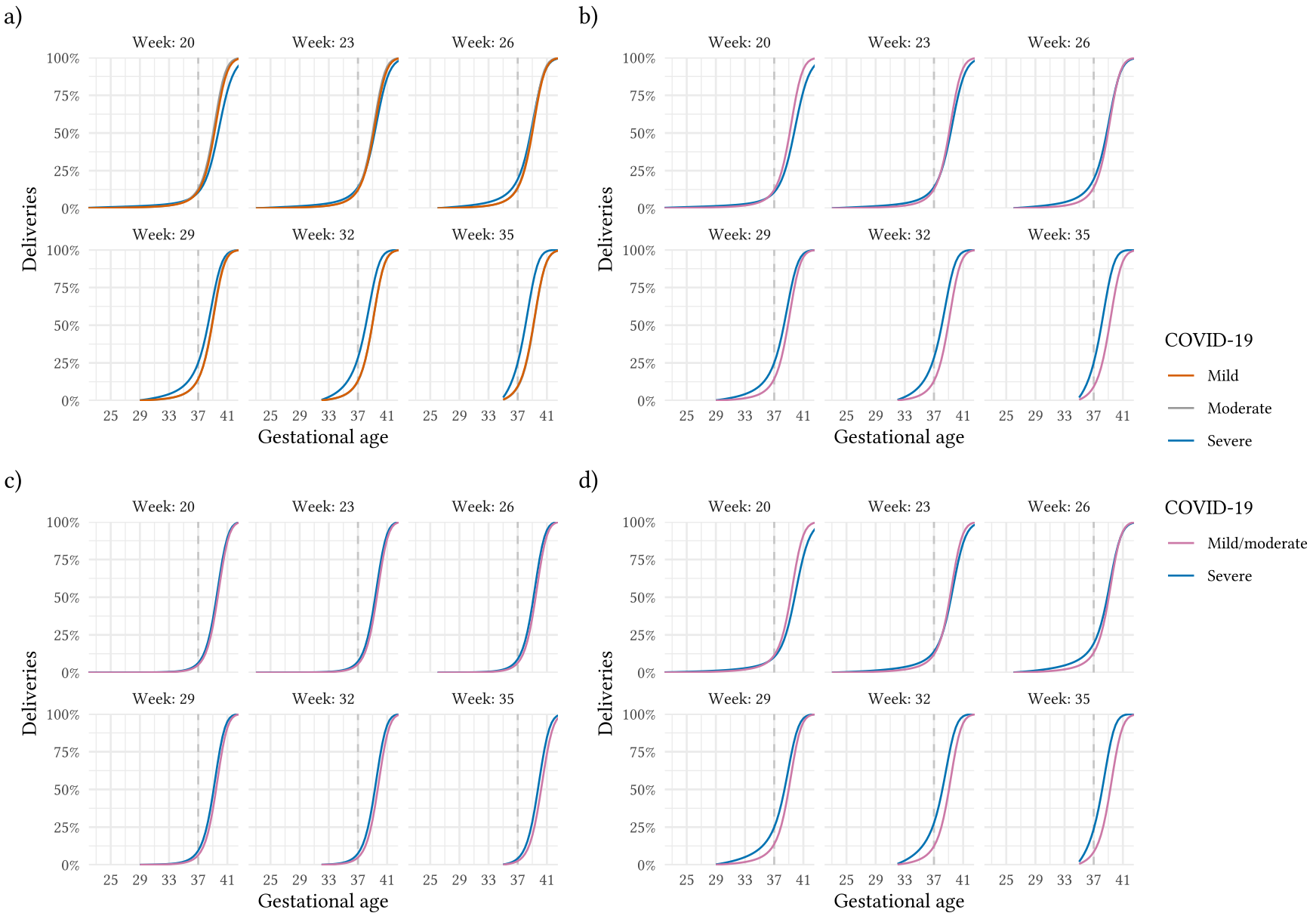

Figure A.5: *Cumulative deliveries across gestation, estimated in various sensitivity analyses (COVID-19 severity): a) Estimating risks separately for mild and moderate severity groups; b) Finer modeling of the discrete-time hazard of delivery; c) Excluding participants who joined after completion of pregnancy; d) Excluding participants who tested negative within two weeks before delivery (i.e., routine delivery testing).*

B

Appendix to Chapter 2

Table B.1: *Number of person-months and deaths contributing to inverse-probability weighted estimates for all-cause mortality under a range of treatment regimes, defined by the threshold of PSA doubling time in days at which treatment was initiated.*

Threshold	Person-months	Deaths
0	64,247	145
60	57,784	134
120	49,088	132
180	41,271	106
240	33,786	85
300	28,953	72
360	25,205	64

Table B.2: *Description of the models used in the two estimation methods. The value of the time-varying covariates used to fit the models was the most recent value, with the expection of PSA, for which the highest value of the past 3 months was used, and PSADT, for which the lowest value of the past 3 months was used.*

Variable	Baseline covariates	Time-varying covariates	Model type	Subset
Parametric g-formula				
any clinic visit	year of relapse, time to relapse, diagnostic risk category, original treatment, age at diagnosis, comorbidities, PSADT at relapse, PSA at relapse	symptoms, any progression, length of progression, on ADT, duration on ADT, interaction term between ADT indicator and duration, duration not on ADT, PSA, PSADT, high PSADT, time since baseline (natural cubic spline with knots at 12, 29, 56), time since last visit (natural cubic spline with knots at 2 and 5 months)	logistic	all observations
PSA	year of relapse, time to relapse, diagnostic risk category, original treatment, age at diagnosis, comorbidities, PSADT at relapse, PSA at relapse	symptoms, any progression, length of progression, PSA, on ADT, duration on ADT, interaction term between ADT indicator and duration, duration not on ADT, PSADT, high PSADT	logistic for PSA = 0; linear in log(PSA)	all visits (subset to those with non-0 PSA for linear model)
PSADT		PSA (2 most recent measurements), time between 2 most recent PSA measurements	direct calculation	all visits
any symptoms	year of relapse, time to relapse, diagnostic risk category, original treatment, age at diagnosis, comorbidities, PSADT at relapse, PSA at relapse	symptoms, any progression, length of progression, PSA, on ADT, duration on ADT, interaction term between ADT indicator and duration, duration not on ADT, PSADT, high PSADT	logistic	all visits

disease progression	year of relapse, time to relapse, diagnostic risk category, original treatment, age at diagnosis, comorbidities, PSADT at relapse, PSA at relapse	symptoms, PSA, on ADT, duration on ADT, interaction term between ADT indicator and duration, duration not on ADT, PSADT, high PSADT	logistic	all observations that have not yet experienced disease progression
ADT initiation	year of relapse, time to relapse, diagnostic risk category, original treatment, age at diagnosis, comorbidities, PSADT at relapse, PSA at relapse	PSA, symptoms, any progression, length of progression, PSADT, high PSADT, time since baseline	logistic	all observations that have not yet initiated ADT
ADT discontinuation	year of relapse, time to relapse, diagnostic risk category, original treatment, age at diagnosis, comorbidities, PSADT at relapse, PSA at relapse	PSA, symptoms, any progression, length of progression, PSADT, high PSADT, duration of ADT	logistic	all observations that are currently on ADT
death	year of relapse, time to relapse, diagnostic risk category, original treatment, age at diagnosis, comorbidities, PSADT at relapse, PSA at relapse	time since last visit, PSA, symptoms, any progression, length of progression, on ADT, duration on ADT, interaction term between ADT indicator and duration, PSADT, high PSADT, duration not on ADT, time since baseline (natural cubic spline with knots at 12, 29, 56)	logistic	all observations
Inverse probability weighting				
ADT initiation	year of relapse, time to relapse, diagnostic risk category, original treatment, PSA at relapse, age at diagnosis, comordidities, PSADT at relapse, PSA at relapse	PSA, months since visit, symptoms, any progression, length of progression, visit this month, PSADT, high PSADT, time since baseline	logistic	all observations that have not yet started ADT

ADT re-initiation	year of relapse, time to relapse, diagnostic risk category, original treatment, age at diagnosis, comordidities, PSADT at relapse, PSA at relapse	PSA, symptoms, any progression, duration of progression, visit this month, PSADT, high PSADT, duration of ADT, duration of no ADT	logistic	all observations that have discontinued ADT
loss to follow-up	year of relapse, time to relapse, diagnostic risk category, original treatment, age at diagnosis, comordidities, PSADT at relapse, PSA at relapse	PSA, months since visit, any progression, length of progression, visit this month, PSADT, high PSADT, duration of ADT, duration of no ADT	logistic	all observations after 24 months
death	'assigned' PSADT threshold (natural cubic spline with knots at 90, 180, 270), year of relapse, time to relapse, diagnostic risk category, original treatment, age at diagnosis, comorbidities, PSADT at relapse, PSA at relapse	time since baseline (natural cubic spline with knots at 20, 36, 69), interaction terms with threshold	logistic	all observations, weighted and possibly replicated

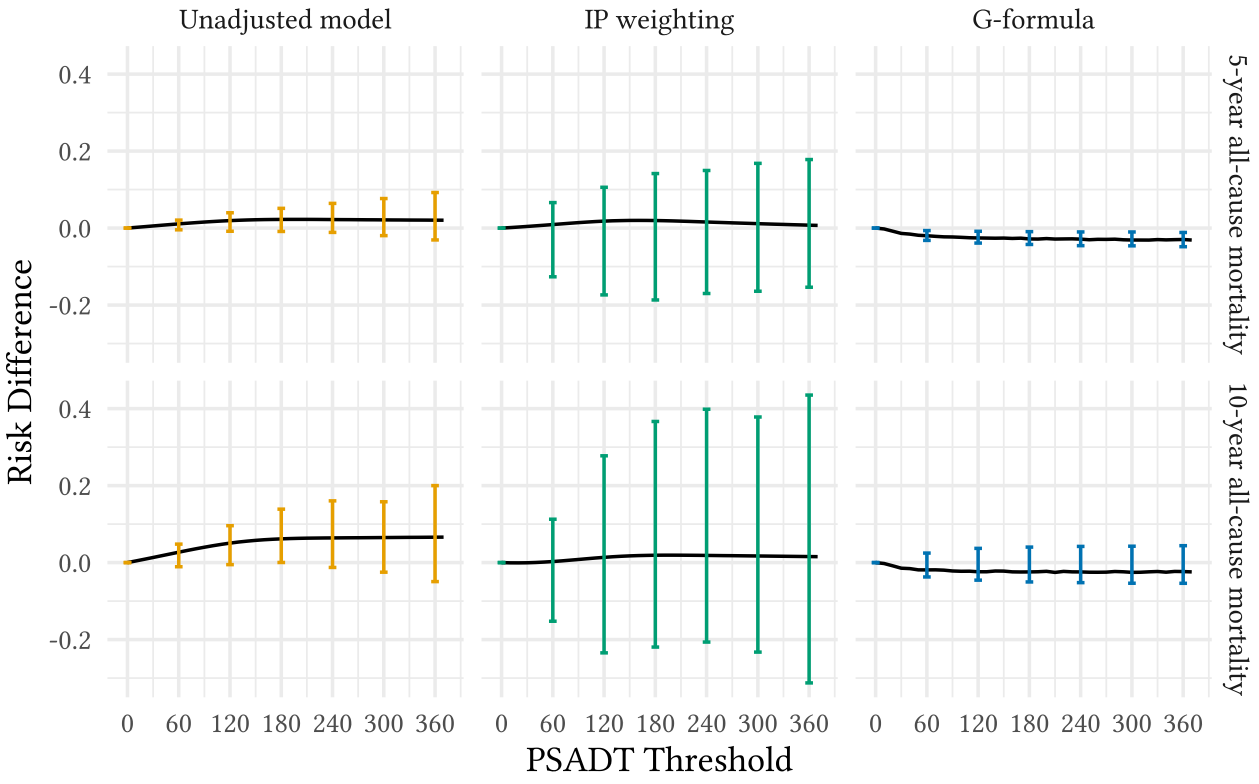

Figure B.1: *Risk differences for 5- and 10-year all-cause mortality across treatment strategies defined by PSADT threshold, compared to the reference threshold of* 0 *(equivalent to never starting treatment based on PSADT). Error bars depict pointwise 95 percent confidence intervals estimated via non-parametric bootstrap, with 1000 replicates.*

Table B.3: *Estimated risk differences for 5- and 10-year all-cause mortality comparing treatment regimes defined by a threshold of PSA doubling time in days at which to initiate treatment, compared to a reference threshold of 0 days. The primary analyses are presented along with an analysis in which treatment is assigned during the grace period according to the observed probability of treatment, instead of uniformly across the grace period as in the main analysis.*

Threshold	IP weighting (primary analyis)	IP weighting (according to observed treatment probabilities)	G-formula (primary analyis)	G-formula (according to observed treatment probabilities)
5-year all-cause mortality				
0	ref.	ref.	ref.	ref.
60	0.01 (-0.13, 0.07)	-0.01	-0.02 (-0.03, -0.01)	-0.02
120	0.02 (-0.17, 0.11)	-0.01	-0.03 (-0.04, -0.01)	-0.02
180	0.02 (-0.19, 0.14)	-0.01	-0.03 (-0.04, -0.01)	-0.02
240	0.02 (-0.17, 0.15)	-0.02	-0.03 (-0.05, -0.01)	-0.03
300	0.01 (-0.16, 0.17)	-0.03	-0.03 (-0.05, -0.01)	-0.03
360	0.01 (-0.15, 0.18)	-0.03	-0.03 (-0.05, -0.01)	-0.03
10-year all-cause mortality				
0	ref.	ref.	ref.	ref.
60	0.00 (-0.15, 0.11)	-0.02	-0.02 (-0.04, 0.02)	-0.02
120	0.01 (-0.23, 0.28)	-0.02	-0.02 (-0.05, 0.04)	-0.03
180	0.02 (-0.22, 0.37)	-0.02	-0.02 (-0.05, 0.04)	-0.03
240	0.02 (-0.21, 0.40)	-0.02	-0.02 (-0.05, 0.04)	-0.03
300	0.02 (-0.23, 0.38)	-0.03	-0.03 (-0.05, 0.04)	-0.03
360	0.02 (-0.31, 0.44)	-0.03	-0.02 (-0.05, 0.04)	-0.03

Table B.4: *Estimated risk differences for 5- and 10-year all-cause mortality comparing treatment regimes defined by a threshold of average PSA doubling time in days at which to initiate treatment, compared to a reference threshold of 0 days.*

Threshold	Unadjusted model	IP weighting	G-formula
5-year all-cause mortality			
0	ref.	ref.	ref.
300	0.00	-0.01	-0.02
600	0.00	-0.03	-0.03
900	-0.01	-0.05	-0.03
1200	-0.01	-0.05	-0.03
1500	-0.01	-0.05	-0.03
1800	-0.01	-0.05	-0.03
10-year all-cause mortality			
0	ref.	ref.	ref.
300	0.00	0.00	-0.01
600	-0.01	-0.03	-0.02
900	-0.01	-0.06	-0.02
1200	0.00	-0.08	-0.02
1500	0.03	-0.08	-0.03
1800	0.07	-0.07	-0.03

C

Appendix to Chapter 3

C.1 A bound for outcome misclassification, selection bias, and unmeasured confounding

Result 1

Let A denote a binary exposure of interest, Y a binary outcome and Y^* the misclassified version, and C measured covariates. Additionally let S be a binary indicator of selection into a study, so that we can collect data only on the subset of the population for which $S = 1$. Finally, assume that there exist U_s and U_c such that $Y \perp\!\!\!\perp S \mid A, C, U_s$ and $Y_a \perp\!\!\!\perp A \mid C, U_c$, but that it is not necessarily true that $Y \perp\!\!\!\perp S \mid A, C$ or $Y_a \perp\!\!\!\perp A \mid C$.

We can estimate a confounded risk ratio observed in the selected population, subject to (potentially differential) outcome misclassification, $\mathrm{RR}_{AY}^{\mathrm{obs}}$, but our inferential goal is a causal risk ratio for the true outcome in the entire population, $\mathrm{RR}_{AY}^{\mathrm{true}}$:

$$\mathrm{RR}_{AY}^{\mathrm{obs}} = \frac{\Pr\left(Y^* = 1 \mid A = 1, S = 1, c\right)}{\Pr\left(Y^* = 1 \mid A = 0, S = 1, c\right)}$$
$$\mathrm{RR}_{AY}^{\mathrm{true}} = \frac{\Pr\left(Y_1 = 1 \mid c\right)}{\Pr\left(Y_0 = 1 \mid c\right)}$$

We have from VanderWeele & Li[96] that, for $\mathrm{RR}_{AY}^{\mathrm{true}} \geq 1$,

$$\mathrm{RR}_{AY}^{\mathrm{obs}} \leq \mathrm{BF}_m \times \frac{\Pr\left(Y = 1 \mid A = 1, S = 1, c\right)}{\Pr\left(Y = 1 \mid A = 0, S = 1, c\right)} \tag{C.1}$$

for

$$\mathrm{BF}_m = \mathrm{RR}_{AY^*|y,S=1} = \max_y \frac{\Pr\left(Y^* = 1 \mid Y = y, A = 1, S = 1, c\right)}{\Pr\left(Y^* = 1 \mid Y = y, A = 0, S = 1, c\right)} \,. \tag{C.2}$$

Then, since we are assuming that $Y \perp\!\!\!\perp S \mid A, C, U_s$, from Smith & VanderWeele[95] we have that

$$\frac{\Pr\left(Y = 1 \mid A = 1, S = 1, c\right)}{\Pr\left(Y = 1 \mid A = 0, S = 1, c\right)} \leq \mathrm{BF}_s \times \frac{\Pr\left(Y = 1 \mid A = 1, c\right)}{\Pr\left(Y = 1 \mid A = 0, c\right)} \tag{C.3}$$

for

$$\mathrm{BF}_s = \frac{\mathrm{RR}_{U_sY|A=1} \times \mathrm{RR}_{SU_s|A=1}}{\mathrm{RR}_{U_sY|A=1} + \mathrm{RR}_{SU_s|A=1} - 1} \times \frac{\mathrm{RR}_{U_sY|A=0} \times \mathrm{RR}_{SU_s|A=0}}{\mathrm{RR}_{U_sY|A=0} + \mathrm{RR}_{SU_s|A=0} - 1}$$

where

$$\begin{aligned}
\mathrm{RR}_{U_sY|A=a} &= \frac{\max_u \Pr\left(Y = 1 \mid A = a, c, U_s = u\right)}{\min_u \Pr\left(Y = 1 \mid A = a, c, U_s = u\right)} \quad \text{for } a = 0, 1 \\
\mathrm{RR}_{SU_s|A=1} &= \max_u \frac{\Pr\left(U_s = u \mid A = 1, S = 1, c\right)}{\Pr\left(U_s = u \mid A = 1, S = 0, c\right)} \\
\mathrm{RR}_{SU_s|A=0} &= \max_u \frac{\Pr\left(U_s = u \mid A = 0, S = 0, c\right)}{\Pr\left(U_s = u \mid A = 0, S = 1, c\right)} \,.
\end{aligned} \tag{C.4}$$

Finally, since we are assuming that $Y_a \perp\!\!\!\perp A \mid C, U_c$ from Ding & VanderWeele[94] we have

$$\frac{\Pr\left(Y = 1 \mid A = 1, c\right)}{\Pr\left(Y = 1 \mid A = 0, c\right)} \leq \mathrm{BF}_c \times \frac{\Pr\left(Y_1 = 1 \mid c\right)}{\Pr\left(Y_0 = 1 \mid c\right)} \tag{C.5}$$

for

$$\mathrm{BF}_c = \frac{\mathrm{RR}_{AU_c} \times \mathrm{RR}_{U_cY}}{\mathrm{RR}_{AU_c} + \mathrm{RR}_{U_cY} - 1} \tag{C.6}$$

where

$$\begin{aligned}
\mathrm{RR}_{AU_c} &= \max_u \frac{\Pr\left(U_c = u \mid A = 1, c\right)}{\Pr\left(U_c = u \mid A = 0, c\right)} \\
\mathrm{RR}_{U_cY} &= \max_a \frac{\max_u \Pr\left(Y = 1 \mid A = a, c, U_c = u\right)}{\min_u \Pr\left(Y = 1 \mid A = a, c, U_c = u\right)} \,.
\end{aligned}$$

Putting together expressions (C.1), (C.3), and (C.5), we have Result 1:

$$\begin{aligned}
\mathrm{RR}_{AY}^{\mathrm{obs}} &\leq \mathrm{BF}_m \times \frac{\Pr\left(Y=1 \mid A=1, S=1, c\right)}{\Pr\left(Y=1 \mid A=0, S=1, c\right)} \\
&\leq \mathrm{BF}_m \times \mathrm{BF}_s \times \frac{\Pr\left(Y=1 \mid A=1, c\right)}{\Pr\left(Y=1 \mid A=0, c\right)} \\
&\leq \mathrm{BF}_m \times \mathrm{BF}_s \times \mathrm{BF}_c \times \frac{\Pr\left(Y_1=1 \mid c\right)}{\Pr\left(Y_0=1 \mid c\right)} \\
&= \mathrm{BF}_m \times \mathrm{BF}_s \times \mathrm{BF}_c \times \mathrm{RR}_{AY}^{\mathrm{true}} \ . \qquad \text{(C.7)}
\end{aligned}$$

An alternative decomposition

Now assume that there exist U_s and U_c such that $Y^* \perp\!\!\!\perp S \mid A, C, U_s$ and $Y_a \perp\!\!\!\perp A \mid C, U_c$. This may be the case if, for example, selection into the study is based on a factor related to the (mis)measured outcome, not the true outcome.

Then we can bound the bias with the same final expression, but some of the parameters within the bias factors are defined slightly differently.

The possible magnitude of selection bias can be defined in terms of the misclassified outcome, so that

$$\mathrm{BF}_s = \frac{\mathrm{RR}_{U_s Y^* \mid A=1} \times \mathrm{RR}_{SU_s \mid A=1}}{\mathrm{RR}_{U_s Y^* \mid A=1} + \mathrm{RR}_{SU_s \mid A=1} - 1} \times \frac{\mathrm{RR}_{U_s Y^* \mid A=0} \times \mathrm{RR}_{SU_s \mid A=0}}{\mathrm{RR}_{U_s Y^* \mid A=0} + \mathrm{RR}_{SU_s \mid A=0} - 1}$$

where

$$\mathrm{RR}_{U_s Y^* \mid A=a} = \frac{\max_u \Pr\left(Y^*=1 \mid A=a, c, U_s=u\right)}{\min_u \Pr\left(Y^*=1 \mid A=a, c, U_s=u\right)} \quad \text{for } a=0,1$$

and $\mathrm{RR}_{SU_s \mid A=1}$ and $\mathrm{RR}_{SU_s \mid A=0}$ are defined as in (C.4) above. Then, the measurement error correction applies to the entire population, so that

$$\mathrm{BF}_m = \mathrm{RR}_{AY^* \mid y} = \max_y \frac{\Pr\left(Y^*=1 \mid Y=y, A=1, c\right)}{\Pr\left(Y^*=1 \mid Y=y, A=0, c\right)} \ .$$

Expression (C.7) now holds with the newly defined BF_s and BF_m.

C.2 A bound for exposure misclassification, selection bias, and unmeasured confounding

Unlike the bound for outcome misclassification, the bound for exposure misclassification from VanderWeele & Li[96] applies to the odds ratio, not the risk ratio, and the sensitivity parameters are also not risk ratios. That is,

$$\frac{\frac{\Pr(Y=1|A^*=1,c)}{\Pr(Y=0|A^*=1,c)}}{\frac{\Pr(Y=1|A^*=0,c)}{\Pr(Y=0|A^*=0,c)}} \leq \mathrm{BF}'_m \times \frac{\frac{\Pr(Y=1|A=1,c)}{\Pr(Y=0|A=1,c)}}{\frac{\Pr(Y=1|A=0,c)}{\Pr(Y=0|A=0,c)}} \tag{C.8}$$

for

$$\mathrm{BF}'_m = \mathrm{OR}_{YA^*|a} = \max\left(\frac{\frac{s'_1}{1-s'_1}}{\frac{s'_0}{1-s'_0}}, \frac{\frac{f'_1}{1-f'_1}}{\frac{f'_0}{1-f'_0}}, \frac{\frac{f'_1}{f'_0}}{\frac{1-s'_1}{1-s'_0}}, \frac{\frac{s'_1}{s'_0}}{\frac{1-f'_1}{1-f'_0}}\right) \tag{C.9}$$

where $s'_y = \Pr(A^* = 1 \mid Y = y, A = 1, c)$ and $f'_y = \Pr(A^* = 1 \mid Y = y, A = 0, c)$. Applying this bound after factoring out selection bias, we would find that we are left with

$$\mathrm{RR}^{\mathrm{obs}}_{AY} \leq \mathrm{BF}'_m \times \mathrm{BF}_s \times \mathrm{BF}_c \times \mathrm{RR}^{\mathrm{true}}_{AY} \times \frac{\Pr(Y = 0 \mid A = 0, c)}{\Pr(Y = 0 \mid A = 1, c)} \times \frac{\Pr(Y = 0 \mid A^* = 1, c)}{\Pr(Y = 0 \mid A^* = 0, c)}$$

for some BF'_m, BF_s, and BF_c, which is not as useful for sensitivity analysis. However, if the outcome is sufficiently rare that $\Pr(Y = 0 \mid \cdot) \approx 1$ in all strata, a simpler bound holds approximately, as we show next.

Again we can define the parameters in the bound in two ways by considering two sets of assumptions.

Result 2

If there exist U_s and U_c such that $Y \perp\!\!\!\perp S \mid A, C, U_s$ and $Y_a \perp\!\!\!\perp A \mid C, U_c$, and if $\Pr(Y = 0 \mid \cdot) \approx 1$, then we have Result 2:

$$\begin{aligned}\mathrm{RR}^{\mathrm{obs}'}_{AY} &= \frac{\Pr(Y = 1 \mid A^* = 1, S = 1, c)}{\Pr(Y = 1 \mid A^* = 0, S = 1, c)} \\ &\lesssim \mathrm{BF}'_m \times \mathrm{BF}_s \times \mathrm{BF}_c \times \mathrm{RR}^{\mathrm{true}}_{AY}\end{aligned}$$

for $\text{BF}'_m = \text{OR}_{YA^*|a,S=1}$ equivalent to the expression (C.9), but with $s'_y = \Pr(A^* = 1 \mid Y = y, A = 1, S = 1, c)$ and $f'_y = \Pr(A^* = 1 \mid Y = y, A = 0, S = 1, c)$; BF_s as defined in (C.4); and BF_c as defined in (C.6).

An alternative decomposition

Alternatively, if $Y \perp\!\!\!\perp S \mid A^*, C, U_s$ and $Y_a \perp\!\!\!\perp A \mid C, U_c$, then the bound holds approximately with

$$\text{BF}_s = \frac{\text{RR}_{U_sY|A^*=1} \times \text{RR}_{SU_s|A^*=1}}{\text{RR}_{U_sY|A^*=1} + \text{RR}_{SU_s|A^*=1} - 1} \times \frac{\text{RR}_{U_sY|A^*=0} \times \text{RR}_{SU_s|A^*=0}}{\text{RR}_{U_sY|A^*=0} + \text{RR}_{SU_s|A^*=0} - 1}$$

where $\text{RR}_{U_sY|A^*=a}$ and $\text{RR}_{SU_s|A^*=0}$ are defined as above, with all A replaced with A^* and Y^* replaced with Y, and with BF'_m as originally defined in expression (C.9).

Interpretation of the exposure misclassification parameters

While all of the sensitivity parameters we have considered thus far are risk ratios, we have seen that those making up the bound for exposure misclassification are not. If, however, the misclassified exposure is sufficiently rare that $\Pr(A^* = 0 \mid \cdot) \approx 1$, then we can interpret the sensitivity parameters as risk ratios:

$$\text{BF}'_m = \text{RR}_{YA^*|a} = \max_a \left(\frac{\Pr(A^* = 1 \mid Y = 1, A = a, c)}{\Pr(A^* = 1 \mid Y = 0, A = a, c)} \right) \quad \text{or}$$
$$\text{BF}'_m = \text{RR}_{YA^*|a,S=1} = \max_a \left(\frac{\Pr(A^* = 1 \mid Y = 1, A = a, S = 1, c)}{\Pr(A^* = 1 \mid Y = 0, A = a, S = 1, c)} \right).$$

Alternatively, if the exposure is not particularly rare, we can interpret the sensitivity parameters as squares of the RR equivalents, using the square-root approximation of the odds ratio.[120]

C.3 Inference in the selected population

Result 3 (Under outcome misclassification)

It may be that our target of inference is the selected population only, so that

$$\mathrm{RR}^{\mathrm{true}}_{AY|S=1} = \frac{\Pr\left(Y_1 = 1 \mid S = 1, c\right)}{\Pr\left(Y_0 = 1 \mid S = 1, c\right)} .$$

In this case we need that assumption $Y_a \perp\!\!\!\perp A \mid S = 1, C, U_c, U_s$: we must simultaneously consider both the factor(s) creating selection bias and the factor(s) creating confounding (which may be one and the same). Let U_{sc} denote the vector (U_s, U_c). Then after factoring out the misclassification bias, we have Result 3:

$$\begin{aligned}
\mathrm{RR}^{\mathrm{obs}}_{AY} &\leq \mathrm{BF}_m \times \frac{\Pr\left(Y = 1 \mid A = 1, S = 1, c\right)}{\Pr\left(Y = 1 \mid A = 0, S = 1, c\right)} \\
&\leq \mathrm{BF}_m \times \mathrm{BF}_{sc} \times \frac{\Pr\left(Y_1 = 1 \mid S = 1, c\right)}{\Pr\left(Y_0 = 1 \mid S = 1, c\right)} \\
&= \mathrm{BF}_m \times \mathrm{BF}_{sc} \times \mathrm{RR}^{\mathrm{true}}_{AY|S=1} \qquad \text{(C.10)}
\end{aligned}$$

for

$$\mathrm{BF}_{sc} = \frac{\mathrm{RR}_{AU_{sc}} \times \mathrm{RR}_{U_{sc}Y}}{\mathrm{RR}_{AU_{sc}} + \mathrm{RR}_{U_{sc}Y} - 1}$$

where

$$\begin{aligned}
\mathrm{RR}_{AU_{sc}} &= \max_u \frac{\Pr\left(U_{sc} = u \mid A = 1, S = 1, c\right)}{\Pr\left(U_{sc} = u \mid A = 0, S = 1, c\right)} \\
\mathrm{RR}_{U_{sc}Y} &= \max_a \frac{\max_u \Pr\left(Y = 1 \mid A = a, S = 1, c, U_{sc} = u\right)}{\min_u \Pr\left(Y = 1 \mid A = a, S = 1, c, U_{sc} = u\right)}
\end{aligned}$$

and BF_m is defined as in (C.2).

Under exposure misclassification

Again we consider the bias due to selection and unmeasured confounding jointly. The bound in expression (C.10) holds with BF'_m constructed with $s'_y = \Pr\left(A^* = 1 \mid Y = y, A = 1, S = 1, c\right)$ and $f'_y = \Pr\left(A^* = 1 \mid Y = y, A = 0, S = 1, c\right)$.

C.4 The multi-bias E-value

The bounds in Results 1, 2, and 3 allow researchers and consumers of research to choose values for bias parameters and investigate their possible effects on an observed risk ratio. Target-adjusted sensitivity analysis, on the other hand, quantifies the strength of bias necessary to shift an observation to another value, often the null value, though others can be used.[121] The E-value for unmeasured confounding is an example of this approach.[102] We can calculate an equivalent value for a combination of biases using the bounds in this article. The E-value for unmeasured confounding refers to a value that can be shown to be sufficient to explain away an observed estimate and that jointly minimizes the maximum of the two sensitivity parameters for unmeasured confounding.[102] Similarly, the multi-bias E-value describes the minimum value that all of the sensitivity parameters for each of the biases would have to take on for a given observed risk ratio to be compatible with a truly null risk ratio. Since the overall bias is monotone increasing in the individual bias parameters, it follows that if any one of the bias parameters is less than the multi-bias E-value, then at least one other parameter would have to be greater than the multi-bias E-value in order to completely explain a result.

Recall that under non-differential misclassification of the exposure, the BF_m factor in the bound is not a risk ratio. If the misclassified exposure is rare, then that parameter can be interpreted as an approximate risk ratio; otherwise, an approximate square root transformation for the odds ratio can be applied so as to approximate the risk ratio.[96] In this way all the parameters that the multi-bias E-value pertains to are on the (approximate) risk ratio scale.

Figure 2 shows the size of the multiple-bias E-value for various combinations of biases and across a range of observed risk ratios. In general, this demonstrates that when there are multiple forms of bias, very little of each type could be sufficient to produce a risk ratio that is within the range we generally see in epidemiologic studies. For example, when the null is true, it is possible to observe a risk ratio of 4 if each of the outcome misclassification ($\text{RR}_{AY^*|y,S=1}$), selec-

tion bias ($RR_{U_sY|A=1}, RR_{U_sY|A=1}, RR_{SU_s|A=1}, RR_{SU_s|A=0}$), and unmeasured confounding (RR_{U_cY}, RR_{AU_c}) parameters is approximately 1.89.

Of course, it is unlikely that each of these sensitivity analysis parameters would be equal to the others, and equal to 1.89. The bounds in this article can be used to assess the bias with a more realistic set of parameters. However, comparing multiple-bias E-values for various combinations of biases may be useful when planning studies to assess where resources should be invested to avoid certain biases, or to assess where a more in-depth bias analysis would be most useful.

Unfortunately, we know of no closed-form solution for this value when we are faced with all three types of bias, but it is easily solved numerically. The expressions to be solved are given in the final column of Table 1. To calculate the analogous multi-bias E-value needed to shift the observed RR_{AY}^{obs} to some risk ratio, RR_{AY}^{true}, other than the null, one can simply replace RR_{AY}^{obs} in the each formula with $RR_{AY}^{obs}/RR_{AY}^{true}$. Also, each formula presupposes that $RR_{AY}^{obs} >= 1$; for apparently protective exposures, the inverse should be taken first.

We will demonstrate interpretation of the multiple bias E-value with respect to our examples, and then briefly describe an R package that can be used to implement the results.

Examples

Recall from the main text that the study of HIV infection in children found $RR_{AY}^{obs} = 6.75$,[100] which we determined was possibly affected by selection bias and unmeasured confounding. The multi-bias E-value for that study, given the assumptions about bias we have made, is 4.64. This tells us that $RR_{U_sY|A=1} = RR_{SU_s|A=1} = RR_{AU_c} = RR_{U_cY} =\geq 4.64$ could suffice to completely explain the observed result, but weaker combined bias would not. If, for example, selection bias were indeed weaker, the strength of the unmeasured confounding parameters would have to be stronger than 4.64 for the observation to be compatible with a truly null effect. Repeating the calculation with the lower limit of the confidence interval, we obtain a multi-bias E-value of 2.73. If all of the parameters were this large, it is possible that the confidence interval would include the null.

The estimate from the vitamins-leukemia study was $\text{RR}^{\text{obs}}_{A^*Y} = 0.51$.[101] After taking the inverse so that$\text{RR}^{\text{obs}}_{A^*Y} = 1/0.51 = 1.96$, we find that the multi-bias E-value for exposure misclassification and unmeasured confounding is 1.35. In order to interpret that number consistently across biases, the multi-bias E-value we have calculated pertains to RR_{AU_c}, RR_{U_cY}, and $\text{RR}_{YA^*|a}$, the latter being the square-root approximation of the $\text{OR}_{YA^*|a}$ term in the bound for exposure misclassification.[96] This allows us to interpret 1.35 as the minimum strength on the risk ratio scale that an unmeasured confounder, or set of confounders, would have to have on the outcome, that would have to relate vitamin use to the confounder, and that the false positive probability or sensitivity for vitamin use would have to be increase by, in order for these biases to explain the entire observed risk ratio. Again, this is simply a heuristic, not something we would expect to be the case; for example, we might expect weaker misclassification but stronger confounding. For the limit of the confidence interval closest to the null, 0.89, if we take inverses, we obtain $1/0.89 = 1.12$ and the multi-bias E-value for this is only 1.06, indicating that whether the true risk ratio is smaller than or greater than 1 is indeed sensitive to relatively small amounts of bias.

Derivation

To form a multiple bias E-value,[102] we can set all of the parameters that make up the terms in the bounds equal to each other, then solve for that value to see what magnitude of bias would result in an $\text{RR}^{\text{obs}}_{AY}$ of at least the value observed, if $\text{RR}^{\text{true}}_{AY} = 1$.

For example, for the bound for outcome misclassification, general selection bias, and unmeasured confounding:

$$\begin{aligned}
\mathrm{RR}_{AY}^{\mathrm{obs}} &\leq \max \mathrm{RR}_{AY^*|y,S=1} \times \frac{\mathrm{RR}_{U_sY|A=1} \times \mathrm{RR}_{SU_s|A=1}}{\mathrm{RR}_{U_sY|A=1} + \mathrm{RR}_{SU_s|A=1} - 1} \times \\
&\quad \frac{\mathrm{RR}_{U_sY|A=0} \times \mathrm{RR}_{SU_s|A=0}}{\mathrm{RR}_{U_sY|A=0} + \mathrm{RR}_{SU_s|A=0} - 1} \times \frac{\mathrm{RR}_{AU_c} \times \mathrm{RR}_{U_cY}}{\mathrm{RR}_{AU_c} + \mathrm{RR}_{U_cY} - 1} \times 1 \\
&= x \times \frac{x^2}{2x-1} \times \frac{x^2}{2x-1} \times \frac{x^2}{2x-1} \\
&= \frac{x^7}{(2x-1)^3} \qquad \text{(C.11)}
\end{aligned}$$

for $x = \mathrm{RR}_{AY^*|y,S=1} = \mathrm{RR}_{U_sY|A=1} = \mathrm{RR}_{SU_s|A=1} = \mathrm{RR}_{U_sY|A=0} = \mathrm{RR}_{SU_s|A=0} = \mathrm{RR}_{AU_c} = \mathrm{RR}_{U_cY}$. To our knowledge, this polynomial has no closed-form solution. However, we can easily solve it numerically.

For example, if $\mathrm{RR}_{AY}^{\mathrm{obs}} = 3$, then $x = 1.71$, meaning that if each of the parameters were at least 1.71, the observed risk ratio could be consistent with a truly null causal risk ratio. If any of the parameters were smaller than 1.71, others would have to be larger if the causal risk ratio were truly null.

We can solve the inequality for any combination of parameters that make up a particular bound in a given situation (e.g., for outcome misclassification and selection bias only, or for exposure misclassification with a rare outcome and unmeasured confounding). When considering exposure misclassification, to calculate a multiple bias E-value, we first must confirm that the outcome is rare. Then, if the misclassified exposure is rare, we can solve equation (C.11) and interpret it with respect to the appropriate parameters; if the exposure is not rare, we can solve

$$\begin{aligned}
\mathrm{RR}_{AY}^{\mathrm{obs}} &\lesssim \mathrm{RR}^2_{YA^*|a,S=1} \times \frac{\mathrm{RR}_{U_sY|A=1} \times \mathrm{RR}_{SU_s|A=1}}{\mathrm{RR}_{U_sY|A=1} + \mathrm{RR}_{SU_s|A=1} - 1} \times \\
&\quad \frac{\mathrm{RR}_{U_sY|A=0} \times \mathrm{RR}_{SU_s|A=0}}{\mathrm{RR}_{U_sY|A=0} + \mathrm{RR}_{SU_s|A=0} - 1} \times \frac{\mathrm{RR}_{AU_c} \times \mathrm{RR}_{U_cY}}{\mathrm{RR}_{AU_c} + \mathrm{RR}_{U_cY} - 1} \times 1 \\
&= x^2 \times \frac{x^2}{2x-1} \times \frac{x^2}{2x-1} \times \frac{x^2}{2x-1} \\
&= \frac{x^8}{(2x-1)^3}
\end{aligned}$$

and interpret with respect to the same parameters.

C.5 Implementation in R

We can use new functions from the R package `EValue`[107] to either calculate the appropriate multiple bias E-value or to calculate a bound for the bias, given proposed parameters. The primary new functions in the package, `multi_bound()` and `multi_evalue()`, accept a set of biases (out of `confounding()`, `selection()`, and `misclassification()`, which take various arguments describing the bias in more detail). The function `multi_bias()` is used to declare those biases. The `multi_bound()` function requires values for the parameters making up the bound for the biases in question. The `multi_evalue()` function requires just a value for the observed risk ratio, and prints a message to the user about the sensitivity parameters it refers to.

We will demonstrate the new package functionality by working through the examples in the main text. We will then show how the new functions can be used to recreate examples from earlier literature as well.

```
library(EValue)
```

Examples from the main text

The `multi_bias()` function takes as arguments one or more of the three bias functions, `confounding()`, `selection()`, and `misclassification()`. They should be listed in the order in which they occur in the data (i.e., does the measurement happen in the sample, or is the sample selected based on mismeasured exposure or outcome values?). Each of `selection()` and `misclassification()` take additional arguments depending on the assumptions and simplifications of a given scenario.

In the HIV example, we were interested in the composite bias due to confounding and selection. We were willing to make the assumption that the outcome is more likely in the selected portion of both exposure groups, so we include the argument `"increased risk"`. (The `"general"` argument is in contrast to `"selected"`, the latter meaning that we are only interested in inference in the selected population. Since `"general"` is the default, we could leave it out.)

```
HIV_biases <- multi_bias(confounding(),
                         selection("general", "increased risk"))
```

Printing the biases prints out the arguments that are required for the `multi_bound()` function for easy copying and pasting into that function.

```
HIV_biases
multi_bound(biases = HIV_biases,
            RRAUc = 2.3, RRUcY = 2.5, RRUsYA1 = 3, RRSUsA1 = 2)
```

```
[1] 2.269737
```

Because the labeling of the arguments is not necessarily intuitive, we might want to confirm which refers to which parameter. We can use the `summary()` function on a object created with the `multi_bias()` function to print more information about the biases.

```
summary(HIV_biases)
```

```
         bias     output argument
1 confounding     RR_AUc    RRAUc
2 confounding     RR_UcY    RRUcY
3   selection RR_UsY|A=1  RRUsYA1
4   selection RR_SUs|A=1  RRSUsA1
```

For easy copying and pasting of the notation we used in this appendix and in the main text, the argument `latex = TRUE` can be used in the summary function to print out an additional column with the parameters in our notation.

To calculate a multi-bias E-value, we must provide the observed effect estimate along with the set of biases. There are two options for doing so. The first is to declare the effect estimate with one of `RR()`, `OR()`, or `HR()`, depending on whether it is a risk, odds, or hazard ratio.

```
multi_evalue(biases = HIV_biases,
             est = OR(6.75, rare = TRUE),
             lo = 2.79, hi = 16.31)
```

```
                        point    lower upper
RR                   6.750000 2.790000 16.31
Multi-bias E-values 4.635703 2.728474    NA
```

The lower and upper bound of the confidence interval are assumed to be on the same scale.

Next we will look at the vitamins-leukemia example from the text. The `misclassification()` bias requires one of either `"outcome"` or `"exposure"`; if exposure misclassification is of interest, the user is also required to specify whether the outcome and/or exposure are sufficiently rare to use a risk ratio approximation for an odds ratio (irrespective of whether the effect estimate is actually on the odds ratio scale).

```
leuk_biases <- multi_bias(confounding(),
                          misclassification("exposure",
                                            rare_outcome = TRUE,
                                            rare_exposure = FALSE))
leuk_biases
```

Again we can calculate the bound and multi-bias E-value as in the text.

```
multi_bound(biases = leuk_biases, RRAUc = 2, RRUcY = 1.22, ORYAa = 1.59)
```

```
[1] 1.747568
```

```
multi_evalue(biases = leuk_biases,
             est = OR(0.51, rare = TRUE),
             lo = 0.3, hi = 0.89)
```

```
                        point lower    upper
RR                   0.510000   0.3 0.890000
Multi-bias E-values 1.351985    NA 1.058404
```

We can easily demonstrate that the E-value is the same whether or not the effect estimate is inverted if the exposure is apparently protective. Also, if we don't want the message about the parameters to print, we can use the argument `verbose = FALSE`.

```
multi_evalue(biases = leuk_biases,
             est = OR(1/0.51, rare = TRUE),
             hi = 1/0.3, lo = 1/0.89,
             verbose = FALSE)
```

```
                       point    lower    upper
RR                  1.960784 1.123596 3.333333
Multi-bias E-values 1.351985 1.058404       NA
```

Finally, we presented a multi-bias E-value for all three biases. We can use the `summary()` function to just print the single value, instead of the matrix of the estimates and confidence limits and E-values for both.

```
summary(multi_evalue(biases = multi_bias(confounding(),
                                         selection("general"),
                                         misclassification("outcome")),
                     est = RR(4)))
```

```
[1] 1.888478
```

Extensions not appearing in the main text

We may want to vary the magnitude of the parameters used to calculate the bounds. We'll use the biases from the HIV example to demonstrate.

```
# original bound
multi_bound(biases = HIV_biases, RRAUc = 2, RRUcY = 2.5,
            RRUsYA1 = 3, RRSUsA1 = 2)
```

```
[1] 2.142857
```

```
# vary RRAUc from 1.25 to 3
sapply(seq(1.25, 3, by = .25), function(RRAUc) {
  multi_bound(biases = HIV_biases, RRAUc = RRAUc,
```

```
              RRUcY = 2.5, RRUsYA1 = 3, RRSUsA1 = 2)
  })
```

```
[1] 1.704545 1.875000 2.019231 2.142857 2.250000 2.343750 2.426471 2.500000
```

```
# vary RRAUc and RRUcY
param_vals <- seq(1.25, 3, by = .25)

params <- expand.grid(RRAUc = param_vals,
                      RRUcY = param_vals)

vals <- mapply(multi_bound,
               RRAUc = params$RRAUc,
               RRUcY = params$RRUcY,
               MoreArgs = list(biases = HIV_biases,
                               RRUsYA1 = 3, RRSUsA1 = 2))
matrix(vals,
  ncol = length(param_vals),
  dimnames = list(param_vals, param_vals)
)
```

```
         1.25      1.5     1.75        2     2.25      2.5     2.75        3
1.25 1.562500 1.607143 1.640625 1.666667 1.687500 1.704545 1.718750 1.730769
1.5  1.607143 1.687500 1.750000 1.800000 1.840909 1.875000 1.903846 1.928571
1.75 1.640625 1.750000 1.837500 1.909091 1.968750 2.019231 2.062500 2.100000
2    1.666667 1.800000 1.909091 2.000000 2.076923 2.142857 2.200000 2.250000
2.25 1.687500 1.840909 1.968750 2.076923 2.169643 2.250000 2.320312 2.382353
2.5  1.704545 1.875000 2.019231 2.142857 2.250000 2.343750 2.426471 2.500000
2.75 1.718750 1.903846 2.062500 2.200000 2.320312 2.426471 2.520833 2.605263
3    1.730769 1.928571 2.100000 2.250000 2.382353 2.500000 2.605263 2.700000
```

Of course, all of the parameters in the bound could be varied, but summarizing the resulting bounds in a simple table or figure becomes more difficult with more than two dimensions.

When calculating a multi-bias E-value, we may also think that the null is unlikely but wish to consider how much bias could have shifted a different true value to the observed value. For

example, in the HIV example, we could calculate a multi-bias E-value for a true risk ratio of 2 rather than the null value of 1:

```
multi_evalue(biases = HIV_biases,
             est = OR(6.75, rare = TRUE),
             lo = 2.79, hi = 16.31,
             true = 2)
```

```
                      point    lower upper
RR                 6.750000 2.790000 16.31
Multi-bias E-values 3.077243 1.643623    NA
```

The multi-bias E-value for the point estimate, 3.08 is of course smaller than the "null" E-value of 4.64, as less bias could have resulted in an OR of 6.75 if the true OR were 2 than would have been necessary to shift it from 1.

The interpretation of the parameters differs depending on the ordering of the selection bias and misclassification. We can see that the parameters expected in the `multi_bound()` function and printed by the `multi_evalue()` function reflect the ordering in which the biases are added to `multi_bias()` (see `output` column).

```
# misclassification occurs in the selected group
summary(
  multi_bias(selection("general"),
             misclassification("exposure", rare_outcome = TRUE))
  )
```

```
                      bias     output argument
1                selection RR_UsY|A=1  RRUsYA1
2                selection RR_SUs|A=1  RRSUsA1
3                selection RR_UsY|A=0  RRUsYA0
4                selection RR_SUs|A=0  RRSUsA0
5 exposure misclassification OR_YA*|a,S   ORYAaS
```

```r
# selection is of misclassified individuals
summary(
  multi_bias(misclassification("exposure", rare_outcome = TRUE),
             selection("general"))
  )
```

```
                       bias      output argument
1                 selection RR_UsY|A*=0  RRUsYA0
2                 selection RR_SUs|A*=1  RRSUsA1
3                 selection RR_UsY|A*=1  RRUsYA1
4                 selection RR_SUs|A*=1  RRSUsA1
5 exposure misclassification    OR_YA*|a    ORYAa
```

When selection bias and confounding are both of interest, but restricting inference to the selected population only is desired, the parameters are shared by the two biases:

```r
summary(
  multi_bias(confounding(),
             selection("selected"),
             misclassification("exposure", rare_outcome = TRUE))
  )
```

```
                       bias     output argument
1  confounding and selection  RR_AUsc|S  RRAUscS
2  confounding and selection  RR_UscY|S  RRUscYS
3 exposure misclassification OR_YA*|a,S   ORYAaS
```

Finally, we can see the expected relationship between the multi-bias bound and the multi-bias E-value.

```r
biases <- multi_bias(confounding(),
                     selection("general", "decreased risk"),
                     misclassification("outcome"))

# calculate bound with those parameters all equal to 2
multi_bound(biases, RRAUc = 2, RRUcY = 2, RRUsYA0 = 2, RRSUsA0 = 2, RRAYyS = 2)
```

```
[1] 3.555556
```

```
# get multi-bias e-value for that value; should be ~2
summary(multi_evalue(biases, est = RR(3.555556)))
```

```
[1] 1.999997
```

Examples from earlier literature

The multi-bias bound and E-value are generalizations of previously published results. To demonstrate, we recreate here some examples from three articles introducing the bound and E-value concept for confounding, selection bias, and differential misclassification.

From *Sensitivity Analysis without Assumptions*, Ding & VanderWeele 2016[94]

```
# example from page 370
biases_ex1 <- confounding()
# specifying parameters in bound
multi_bound(biases = biases_ex1, RRAUc = 2, RRUcY = 2)
```

```
[1] 1.333333
```

```
# Table 1, page 371
# consider all possible combinations for bound
param_vals <- c(1.3, 1.5, 1.8, 2, 2.5, 3, 3.5, 4, 5, 6, 8, 10)
params <- expand.grid(RRAUc = param_vals,
                      RRUcY = param_vals)
table1_vals <- mapply(multi_bound, RRAUc = params$RRAUc, RRUcY = params$RRUcY,
                      MoreArgs = list(biases = biases_ex1))
table1 <- matrix(table1_vals,
  ncol = length(param_vals),
  dimnames = list(param_vals, param_vals)
)
round(table1, 2)
```

```
     1.3  1.5  1.8    2  2.5    3  3.5    4    5    6    8   10
1.3 1.06 1.08 1.11 1.13 1.16 1.18 1.20 1.21 1.23 1.24 1.25 1.26
1.5 1.08 1.12 1.17 1.20 1.25 1.29 1.31 1.33 1.36 1.38 1.41 1.43
1.8 1.11 1.17 1.25 1.29 1.36 1.42 1.47 1.50 1.55 1.59 1.64 1.67
2   1.13 1.20 1.29 1.33 1.43 1.50 1.56 1.60 1.67 1.71 1.78 1.82
2.5 1.16 1.25 1.36 1.43 1.56 1.67 1.75 1.82 1.92 2.00 2.11 2.17
3   1.18 1.29 1.42 1.50 1.67 1.80 1.91 2.00 2.14 2.25 2.40 2.50
3.5 1.20 1.31 1.47 1.56 1.75 1.91 2.04 2.15 2.33 2.47 2.67 2.80
4   1.21 1.33 1.50 1.60 1.82 2.00 2.15 2.29 2.50 2.67 2.91 3.08
5   1.23 1.36 1.55 1.67 1.92 2.14 2.33 2.50 2.78 3.00 3.33 3.57
6   1.24 1.38 1.59 1.71 2.00 2.25 2.47 2.67 3.00 3.27 3.69 4.00
8   1.25 1.41 1.64 1.78 2.11 2.40 2.67 2.91 3.33 3.69 4.27 4.71
10  1.26 1.43 1.67 1.82 2.17 2.50 2.80 3.08 3.57 4.00 4.71 5.26
```

```
# reduce an observed RR of 2.5 to true value of 1.5, page 371
summary(multi_evalue(biases = confounding(), est = RR(2.5), true = 1.5))
```

```
[1] 2.720763
```

```
# smoking and lung cancer e-value, page 373
summary(multi_evalue(biases = confounding(), est = RR(10.73)))
```

```
[1] 20.94777
```

From *Bounding bias due to selection*, Smith & VanderWeele, 2019[95]

```
biases_ex2 <- selection("general")

# result 1A example
multi_bound(biases = biases_ex2,
            RRUsYA1 = 2, RRSUsA1 = 1.7, RRUsYA0 = 2, RRSUsA0 = 1.5)
```

```
[1] 1.511111
```

```r
# result 1B example
multi_evalue(biases = biases_ex2, est = OR(73.1, rare = TRUE), lo = 13.0)
```

```
                       point     lower upper
RR                   73.10000 13.000000    NA
Multi-bias E-values 16.58415  6.670587    NA
```

```r
# result 4B example
summary(multi_evalue(biases = selection("general", "S = U", "increased risk"),
             est = OR(5.2, rare = TRUE)))
```

```
[1] 5.2
```

```r
# result 5B example
multi_evalue(biases = selection("selected"),
             est = OR(1.5, rare = TRUE), lo = 1.22)
```

```
                       point    lower upper
RR                   1.500000 1.220000    NA
Multi-bias E-values 2.366025 1.738081    NA
```

From *Simple Sensitivity Analysis for Differential Measurement Error*, VanderWeele & Li 2019[96]

```r
biases_ex3 <- misclassification("exposure",
                                rare_outcome = TRUE, rare_exposure = TRUE)
multi_evalue(biases = biases_ex3, est = OR(1.51, rare = TRUE), lo = 1.03)
```

```
                    point lower upper
RR                   1.51  1.03    NA
Multi-bias E-values  1.51  1.03    NA
```

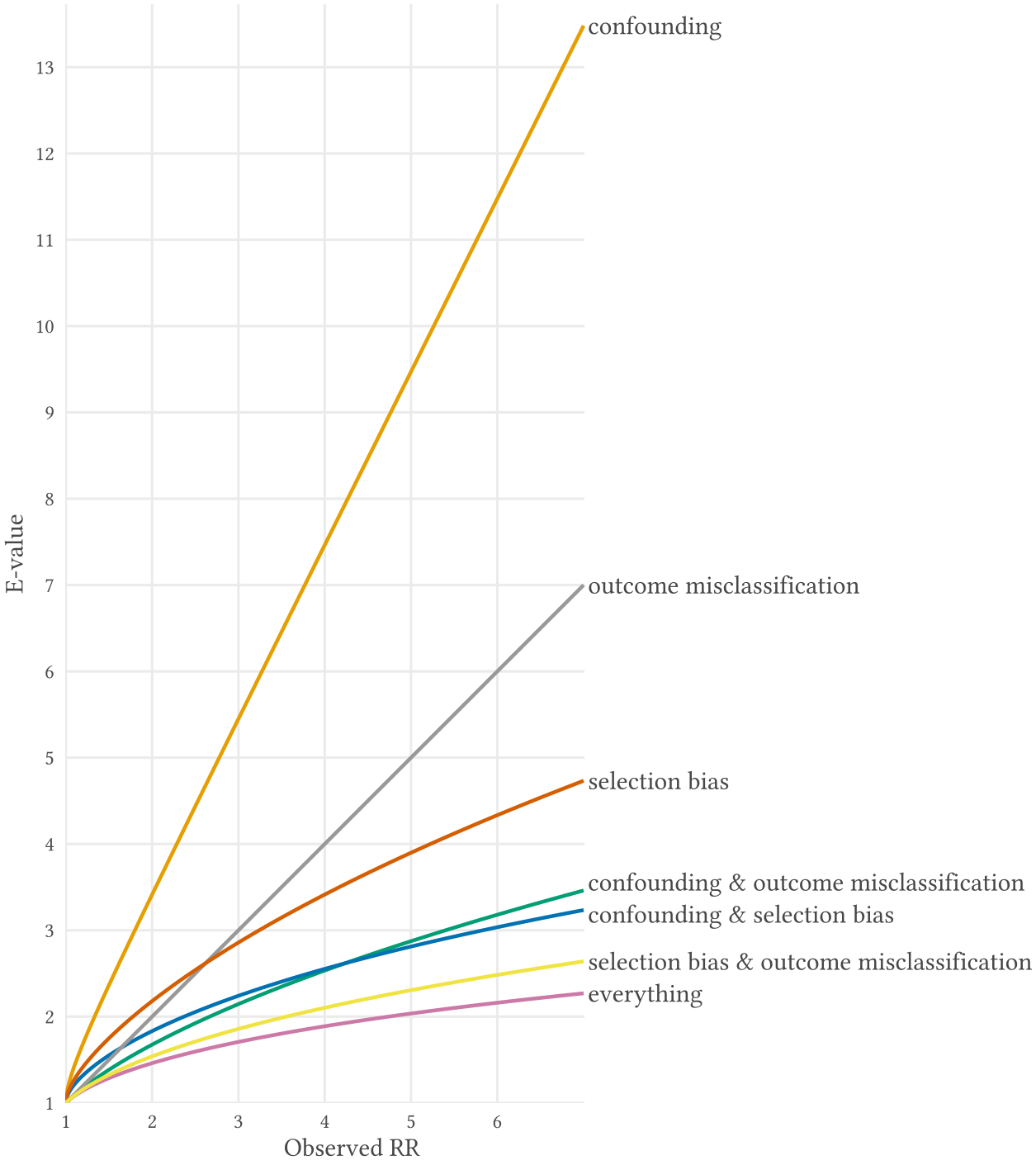

Figure C.1: *Multi-bias E-values for various combinations of biases and for observed risk ratios ranging from 1 to 7.*

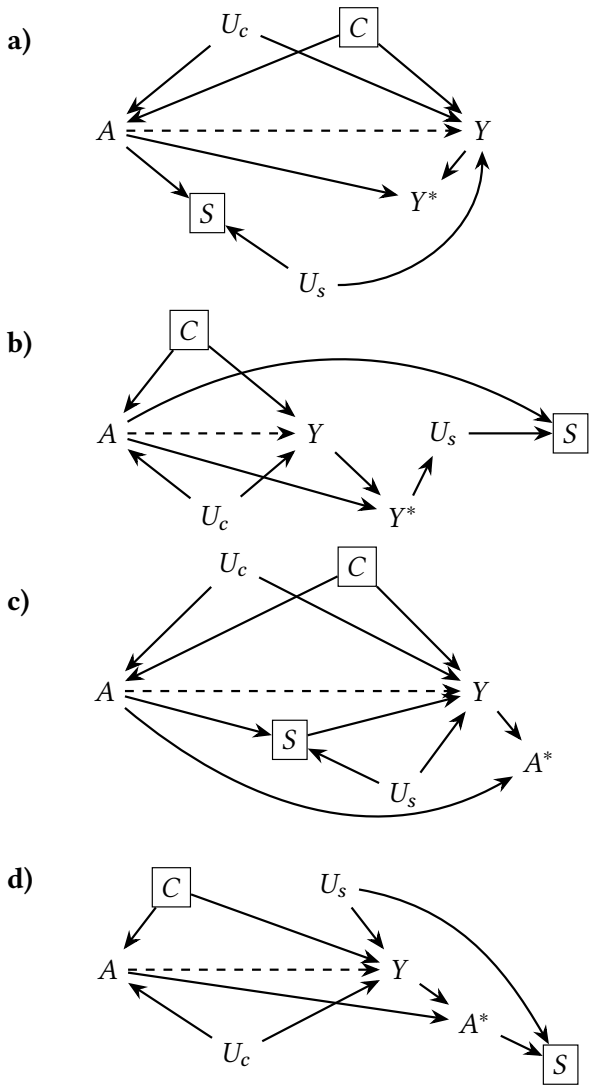

Figure C.2: *Additional directed acyclic graphs depicting multiple biases. These examples show how various combinations of biases can be represented by directed acyclic graphs, and the independence assumptions that are implied.* ***a)*** *This DAG depicts unmeasured confounding (through U_c), selection bias (through U_s), and differential misclassification of the outcome (due to the $A \rightarrow Y^*$ edge). The assumptions $Y \perp\!\!\!\perp S \mid A, C, U_s$ and $Y_a \perp\!\!\!\perp A \mid C, U_c$ are met. This implies that we can apply the outcome misclassification bound, then the selection bias bound, then the unmeasured confounding bound for inference in the total population.* ***b)*** *This DAG depicts unmeasured confounding (through U_c), selection bias (through U_s), and differential misclassification of the outcome (due to the $A \rightarrow Y^*$ edge). The assumptions $Y^* \perp\!\!\!\perp S \mid A, C, U_s$ and $Y_a \perp\!\!\!\perp A \mid C, U_c$ are met. This implies that we can apply the selection bias bound, then the outcome misclassification bound, then the unmeasured confounding bound for inference in the total population.* ***c)*** *This DAG depicts unmeasured confounding (through U_c), selection bias (through U_s), and differential misclassification of the exposure (due to the $Y \rightarrow A^*$ edge). The assumption $Y_a \perp\!\!\!\perp A \mid S = 1, C, U_s, U_c$ is met. This implies that we can apply the the exposure misclassification bound, then the joint bound for selection bias and unmeasured confounding for inference in the selected population.* ***d)*** *This DAG depicts unmeasured confounding (through U_c), selection bias (through U_s), and differential misclassification of the exposure (due to the $Y \rightarrow A^*$ edge). The assumptions $Y \perp\!\!\!\perp S \mid A^*, C, U_s$ and $Y_a \perp\!\!\!\perp A \mid C, U_c$ are met. This implies that we can apply the selection bias bound, then the exposure misclassification bound, then the unmeasured confounding bound for inference in the total population.*

References

1. Yang J, Zheng Y, Gou X, et al. Prevalence of comorbidities and its effects in coronavirus disease 2019 patients: A systematic review and meta-analysis. *Int J Infect Dis.* 2020;94:91-95. doi:10.1016/j.ijid.2020.03.017

2. Zheng Z, Peng F, Xu B, et al. Risk factors of critical & mortal COVID-19 cases: A systematic literature review and meta-analysis. *J Infect.* 2020;81(568):16-25. doi:10.1016/j.jinf.2020.04.021

3. Della Gatta AN, Rizzo R, Pilu G, Simonazzi G. Coronavirus disease 2019 during pregnancy: A systematic review of reported cases. *Am J Obstet Gynecol.* 2020;223(1):36-41. doi:10.1016/j.ajog.2020.04.013

4. Zaigham M, Andersson O. Maternal and perinatal outcomes with COVID-19: A systematic review of 108 pregnancies. *Acta Obstet Gynecol Scand.* 2020;99(7):823-829. doi:10.1111/aogs.13867

5. Sentilhes L, De Marcillac F, Jouffrieau C, et al. COVID-19 in pregnancy was associated with maternal morbidity and preterm birth. *Am J Obstet Gynecol.* 2020;223(6):914.e1-914.e15. doi:10.1016/j.ajog.2020.06.022

6. Barbero P, Mugüerza L, Herraiz I, et al. SARS-CoV-2 in pregnancy: Characteristics and outcomes of hospitalized and non-hospitalized women due to COVID-19. *J Matern Fetal Neonatal Med.* Published online July 20, 2020:1-7. doi:10.1080/14767058.2020.1793320

7. Romagano MP, Guerrero K, Spillane N, et al. Perinatal outcomes in critically ill pregnant women with coronavirus disease 2019. *Am J Obstet Gynecol MFM.* 2020;2(3):100151. doi:10.1016/j.ajogmf.2020.100151

8. Matok I, Azoulay L, Yin H, Suissa S. Immortal time bias in observational studies of drug effects in pregnancy. *Birt Defects Res A Clin Mol Teratol.* 2014;100(9):658-662. doi:10.1002/bdra.23271

9. National Institutes of Health. Clinical Spectrum of SARS-CoV-2 Infection. Published December 17, 2020. Accessed February 19, 2021. https://www.covid19treatmentguidelines.nih.gov/overview/clinical-spectrum/

10. Zambrano LD, Ellington S, Strid P, et al. Update: Characteristics of symptomatic women of reproductive age with laboratory-confirmed SARS-CoV-2 infection by pregnancy status United States, January 22-October 3, 2020. *MMWR Morb Mortal Wkly Rep.* 2020;69:1641-1647. doi:10.15585/mmwr.mm6944e3

11. Suissa S. The Case-Time-Control Design. *Epidemiology.* 1995;6(3):248-253.

12. Suissa S. The Case-Time-Control Design: Further Assumptions and Conditions. *Epidemiology.* 1998;9(4):441-445.

13. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. *Lancet.* 2008;371(9606):75-84. doi:10.1016/S0140-6736(08)60074-4

14. Romero R, Espinoza J, Kusanovic JP, et al. The preterm parturition syndrome. *BJOG Int J Obstet Gynaecol.* 2006;113(s3):17-42. doi:10.1111/j.1471-0528.2006.01120.x

15. Jamieson DJ, Theiler RN, Rasmussen SA. Emerging infections and pregnancy. *Emerg Infect Dis.* 2006;12(11):1638-1643. doi:10.3201/eid1211.060152

16. Kourtis AP, Read JS, Jamieson DJ. Pregnancy and infection. *N Engl J Med.* 2014;370(23):2211-2218. doi:10.1056/NEJMra1213566

17. Siston AM, Rasmussen SA, Honein MA, et al. Pandemic 2009 influenza A(H1N1) virus illness among pregnant women in the United States. *JAMA.* 2010;303(15):1517-1525. doi:10.1001/jama.2010.479

18. Centers for Disease Control and Prevention (CDC). Maternal and infant outcomes among severely ill pregnant and postpartum women with 2009 pandemic influenza A (H1N1)–United States, April 2009-August 2010. *MMWR Morb Mortal Wkly Rep.* 2011;60(35):1193-1196.

19. Doyle TJ, Goodin K, Hamilton JJ. Maternal and neonatal outcomes among pregnant women with 2009 pandemic influenza A(H1N1) illness in Florida, 2009-2010: A population-based cohort study. *PLoS ONE.* 2013;8(10):2009-2010. doi:10.1371/journal.pone.0079040

20. Meijer WJ, Van Noortwijk AGA, Bruinse HW, Wensing AMJ. Influenza virus infection in pregnancy: A review. *Acta Obstet Gynecol Scand.* 2015;94(8):797-819. doi:10.1111/aogs.12680

21. Wong SF, Chow KM, Leung TN, et al. Pregnancy and perinatal outcomes of women with severe acute respiratory syndrome. *Am J Obstet Gynecol.* 2004;191(1):292-297. doi:10.1016/j.ajog.2003.11.019

22. de Souza Silva GA, da Silva SP, da Costa MAS, et al. SARS-CoV, MERS-CoV and SARS-CoV-2 infections in pregnancy and fetal development. *J Gynecol Obstet Hum Reprod.* 2020;49(10):101846. doi:10.1016/j.jogoh.2020.101846

23. Alfaraj SH, Al-Tawfiq JA, Memish ZA. Middle East Respiratory Syndrome Coronavirus (MERS-CoV) infection during pregnancy: Report of two cases & review of the literature. *J Microbiol Immunol Infect.* 2019;52(3):501-503. doi:10.1016/j.jmii.2018.04.005

24. Allotey J, Stallings E, Bonet M, et al. Clinical manifestations, risk factors, and maternal and perinatal outcomes of coronavirus disease 2019 in pregnancy: Living systematic review and meta-analysis. *BMJ.* 2020;370:m3320. doi:10.1136/bmj.m3320

25. Di Toro F, Gjoka M, Di Lorenzo G, et al. Impact of COVID-19 on maternal and neonatal outcomes: A systematic review and meta-analysis. *Clin Microbiol Infect.* 2021;27(1):36-46. doi:10.1016/j.cmi.2020.10.007

26. Matar R, Alrahmani L, Monzer N, et al. Clinical presentation and outcomes of pregnant women with COVID-19: A systematic review and meta-analysis. *Clin Infect Dis.* 2021;72(3):521-533. doi:10.1093/cid/ciaa828

27. Dubey P, Reddy SY, Manuel S, Dwivedi AK. Maternal and neonatal characteristics and outcomes among COVID-19 infected women: An updated systematic review and meta-analysis. *Eur J Obstet Gynecol Reprod Biol.* 2020;252:490-501. doi:10.1016/j.ejogrb.2020.07.034

28. Yee J, Kim W, Han JM, et al. Clinical manifestations and perinatal outcomes of pregnant women with COVID-19: A systematic review and meta-analysis. *Sci Rep.* 2020;10(1):18126. doi:10.1038/s41598-020-75096-4

29. Turan O, Hakim A, Dashraath P, Jeslyn WJL, Wright A, Abdul-Kadir R. Clinical characteristics, prognostic factors, and maternal and neonatal outcomes of SARS-CoV-2 infection among hospitalized pregnant women: A systematic review. *Int J Gynecol Obstet.* 2020;151(1):7-16. doi:10.1002/ijgo.13329

30. Vergara-Merino L, Meza N, Couve-Pérez C, et al. Maternal and perinatal outcomes related to COVID-19 and pregnancy: Overview of systematic reviews. *Acta Obstet Gynecol Scand.* Published online 2021. doi:10.1111/aogs.14118

31. Li N, Han L, Peng M, et al. Maternal and neonatal outcomes of pregnant women with coronavirus disease 2019 (COVID-19) pneumonia: A case-control study. *Clin Infect Dis.* 2020;71(16):2035-2041. doi:10.1093/cid/ciaa352

32. Liao J, He X, Gong Q, Yang L, Zhou C, Li J. Analysis of vaginal delivery outcomes among pregnant women in Wuhan, China during the COVID-19 pandemic. *Int J Gynecol Obstet.* 2020;150(1):53-57. doi:10.1002/ijgo.13188

33. Ahlberg M, Neovius M, Saltvedt S, et al. Association of SARS-CoV-2 test status and pregnancy outcomes. *JAMA*. 2020;324(17):1782-1785. doi:10.1001/jama.2020.19124

34. Flaherman VJ, Afshar Y, Boscardin J, et al. Infant outcomes following maternal infection with SARS-CoV-2: First report from the PRIORITY study. *Clin Infect Dis*. Published online September 18, 2020:ciaa1411. doi:10.1093/cid/ciaa1411

35. Woodworth KR. Birth and infant outcomes following laboratory-confirmed SARS-CoV-2 infection in pregnancy SET-NET, 16 jurisdictions, March 29-October 14, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69(44):1635-1640. doi:10.15585/mmwr.mm6944e2

36. Engjom H, Aabakke AJ, Klungsøyr K, et al. COVID-19 in pregnancy characteristics and outcomes of pregnant women admitted to hospital because of SARS-CoV-2 infection in the Nordic countries. *medRxiv*. Published online February 9, 2021:2021.02.05.21250672. doi:10.1101/2021.02.05.21250672

37. Artymuk NV, Belokrinitskaya TE, Filippov OS, Frolova NI, Surina MN. Perinatal outcomes in pregnant women with COVID-19 in Siberia and the Russian Far East. *J Matern Fetal Neonatal Med*. Published online February 2, 2021. doi:10.1080/14767058.2021.1881954

38. Khoury R, Bernstein PS, Debolt C, et al. Characteristics and outcomes of 241 births to women with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection at five New York City medical centers. *Obstet Gynecol*. 2020;136(2):273-282. doi:10.1097/AOG.0000000000004025

39. Pierce-Williams RAM, Burd J, Felder L, et al. Clinical course of severe and critical coronavirus disease 2019 in hospitalized pregnancies: A United States cohort study. *Am J Obstet Gynecol MFM*. 2020;2(3):100134. doi:10.1016/j.ajogmf.2020.100134

40. Jenabi E, Bashirian S, Khazaei S, et al. Pregnancy outcomes among symptomatic and asymptomatic women infected with COVID-19 in the west of Iran: A case-control study. *J Matern Fetal Neonatal Med*. Published online December 15, 2020:1-3. doi:10.1080/14767058.2020.1861599

41. Panagiotakopoulos L, Myers TR, Gee J, et al. SARS-CoV-2 infection among hospitalized pregnant women: Reasons for admission and pregnancy characteristics eight U.S. Health care centers, March 1-May 30, 2020. *MMWR Morb Mortal Wkly Rep*. 2020;69:5. doi:10.15585/mmwr.mm6938e2

42. Kayem G, Lecarpentier E, Deruelle P, et al. A snapshot of the Covid-19 pandemic among pregnant women in France. *J Gynecol Obstet Hum Reprod*. 2020;49(7):101826. doi:10.1016/j.jogoh.2020.101826

43. Metz TD, Clifton RG, Hughes BL, et al. Disease Severity and Perinatal Outcomes of Pregnant Patients With Coronavirus Disease 2019 (COVID-19). *Obstet Gynecol*. 2021;137(4):571-580. doi:10.1097/AOG.0000000000004339

44. 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

45. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: An application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology.* 2008;19(6):766-779. doi:10.1097/EDE.0b013e3181875e61

46. Tannen RL, Weiner MG, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: Comparison of database and randomised controlled trial findings. *BMJ Online.* 2009;338(7691):395-399. doi:10.1136/bmj.b81

47. Danaei G, Tavakkoli M, Hernán MA. Bias in observational studies of prevalent users: Lessons for comparative effectiveness research from a meta-analysis of statins. *Am J Epidemiol.* 2012;175(4):250-262. doi:10.1093/aje/kwr301

48. Petito LC, García-Albéniz X, Logan RW, et al. Estimates of Overall Survival in Patients With Cancer Receiving Different Treatment Regimens. *JAMA Netw Open.* 2020;3(3):1-13. doi:10.1001/jamanetworkopen.2020.0452

49. Lodi S, Phillips A, Lundgren J, et al. Effect Estimates in Randomized Trials and Observational Studies: Comparing Apples with Apples. *Am J Epidemiol.* 2019;188(8):1569-1577. doi:10.1093/aje/kwz100

50. Artibani W, Porcaro AB, De Marco V, Cerruto MA, Siracusano S. Management of Biochemical Recurrence after Primary Curative Treatment for Prostate Cancer: A Review. *Urol Int.* 2018;100(3):251-262. doi:10.1159/000481438

51. Higano CS. Intermittent Versus Continuous Androgen Deprivation Therapy. *J Natl Compr Canc Netw.* 2014;12(5):727-733. doi:10.6004/jnccn.2014.0074

52. Nguyen PL, Alibhai SMH, Basaria S, et al. Adverse effects of androgen deprivation therapy and strategies to mitigate them. *Eur Urol.* 2015;67(5):825-836. doi:10.1016/j.eururo.2014.07.010

53. Orellana L, Rotnitzky A, Robins JM. Dynamic Regime Marginal Structural Mean Models for Estimation of Optimal Dynamic Treatment Regimes, Part I: Main Content. *Int J Biostat.* 2010;6(2):44398-44399. doi:10.2202/1557-4679.1200

54. Cain LE, Robins JM, Lanoy E, Logan R, Costagliola D, Hernán MA. When to Start Treatment? A Systematic Approach to the Comparison of Dynamic Regimes Using Observational Data. *Int J Biostat.* 2010;6(2):1-42. doi:10.2202/1557-4679.1212

55. Robins J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chronic Dis.* 1987;40(Suppl. 2):139S-161S. doi:10.1016/S0021-9681(87)80018-8

56. Young JG, Cain LE, Robins JM, O'Reilly EJ, Hernán MA. Comparative effectiveness of dynamic treatment regimes: An application of the parametric g-formula. *Stat Biosci.* 2011;3(1):119-143. doi:10.1007/s12561-011-9040-7

57. Garcia-Albeniz X, Chan JM, Paciorek A, et al. Immediate versus deferred initiation of androgen deprivation therapy in prostate cancer patients with PSA-only relapse. An observational follow-up study. *Eur J Cancer.* 2015;51(7):817-824. doi:10.1016/j.ejca.2015.03.003

58. Duchesne GM, Woo HH, Bassett JK, et al. Timing of androgen-deprivation therapy in patients with prostate cancer with a rising PSA (TROG 03.06 and VCOG PR 01-03 [TOAD]): A randomised, multicentre, non-blinded, phase 3 trial. *Lancet Oncol.* 2016;17(6):727-737. doi:10.1016/S1470-2045(16)00107-8

59. Loblaw A, Bassett J, D'Este C, et al. Timing of androgen deprivation therapy for prostate cancer patients after radiation: Planned combined analysis of two randomized phase 3 trials. *J Clin Oncol.* 2018;36(15_suppl):5018-5018. doi:10.1200/jco.2018.36.15_suppl.5018

60. Klayton TL, Ruth K, Buyyounouski MK, et al. Prostate-specific antigen doubling time predicts the development of distant metastases for patients who fail 3-dimensional conformal radiotherapy or intensity modulated radiation therapy using the Phoenix definition. *Pract Radiat Oncol.* 2011;1(4):235-242. doi:10.1016/j.prro.2011.02.003

61. Hernán MA, Robins JM. Per-Protocol Analyses of Pragmatic Trials. *N Engl J Med.* 2017;377(14):1391-1398. doi:10.1056/NEJMsm1605385

62. Pound CR, Partin AW, Eisenberger MA, Chan DW, Pearson JD, Walsh PC. Natural History of Progression After PSA. *J Am Med Assoc.* 1999;281(17):1591-1597. doi:10.1001/jama.281.17.1591

63. Robins JM, Finkelstein DM. Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring Weighted (IPCW) Log-Rank Tests. *Biometrics.* 2000;56(3):779-788. doi:10.1111/j.0006-341X.2000.00779.x

64. Lubeck DP, Litwin MS, Henning JM, et al. The CaPSURE database: A methodology for clinical practice and research in prostate cancer. *Urology.* 1996;48(5):773-777. doi:10.1016/S0090-4295(96)00226-9

65. Cooperberg MR, Broering JM, Litwin MS, et al. The contemporary management of prostate cancer in the United States: Lessons from the Cancer of the Prostate Strategic Urologic Research Endeavor (CaPSURE), a national disease registry. *J Urol.* 2004;171(4):1393-1401. doi:10.1097/01.ju.0000107247.81471.06

66. Studer UE, Collette L, Whelan P, et al. Using PSA to Guide Timing of Androgen Deprivation in Patients with T0-4 N0-2 M0 Prostate Cancer not Suitable for Local Curative Treatment (EORTC 30891). *Eur Urol.* 2008;53(5):941-949. doi:10.1016/j.eururo.2007.12.032

67. McGrath S, Lin V, Zhang Z, et al. gfoRmula: An R Package for Estimating the Effects of Sustained Treatment Strategies via the Parametric g-formula. *Patterns*. 2020;1(3):100008. doi:10.1016/j.patter.2020.100008

68. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2020.

69. Botrel TEA, Clark O, dos Reis RB, et al. Intermittent versus continuous androgen deprivation for locally advanced, recurrent or metastatic prostate cancer: A systematic review and meta-analysis. *BMC Urology*. 2014;14(1):9. doi:10.1186/1471-2490-14-9

70. NCCN Clinical Practice Guidelines in Oncology. *Prostate Cancer*. National Comprehensive Cancer Network; 2018.

71. Dahabreh IJ, Robins JM, Hernán MA. Benchmarking Observational Methods by Comparing Randomized Trials and Their Emulations. *Epidemiology*. 2020;31(5):614-619. doi:10.1097/EDE.0000000000001231

72. Ioannidis JPA. Limitations are not properly acknowledged in the scientific literature. *J Clin Epidemiol*. 2007;60(4):324-329. doi:10.1016/j.jclinepi.2006.09.011

73. Phillips CV. Quantifying and reporting uncertainty from systematic errors. *Epidemiology*. 2003;14(4):459-466. doi:10.1097/01.ede.0000072106.65262.ae

74. Lash TL. Heuristic thinking and inference from observational epidemiology. *Epidemiology*. 2007;18(1):67-72. doi:10.1097/01.ede.0000249522.75868.16

75. Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL. Smoking and lung cancer: Recent evidence and a discussion of some questions. *J Natl Cancer Inst*. 1959;22:173-203. doi:10.1093/ije/dyp289

76. Bross IDJ. Spurious effects from an extraneous variable. *J Chronic Dis*. 1966;19(6):637-647. doi:10.1016/0021-9681(66)90062-2

77. Bross IDJ. Pertinency of an extraneous variable. *J Chronic Dis*. 1967;20(7):487-495. doi:10.1016/0021-9681(67)90080-X

78. Schlesselman JJ. Assessing effects of confounding variables. *Am J Epidemiol*. 1978;108(1):3-8. doi:10.1093/oxfordjournals.aje.a112581

79. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55. doi:10.2307/2335942

80. 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. doi:10.1093/oxfordjournals.aje.a112408

81. Barron BA. The effects of misclassification on the estimation of relative risk. *Biometrics*. 1977;33(2):414-418. doi:10.2307/2529795

82. Greenland S, Neutra R. An analysis of detection bias and proposed corrections in the study of estrogens and endometrial cancer. *J Chronic Dis*. 1981;34(9-10):433-438. doi:10.1016/0021-9681(81)90002-3

83. Greenland S, Ericson C, Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair studies. *Int J Epidemiol*. 1983;12(1):93-97. doi:10.1093/ije/12.1.93

84. Lash TL, Silliman RA. A sensitivity analysis to separate bias due to confounding from bias due to predicting misclassification by a variable that does both. *Epidemiology*. 2000;11(5):544-549. doi:10.1097/00001648-200009000-00010

85. Greenland S. The impact of prior distributions for uncontrolled confounding and response bias: A case study of the relation of wire codes and magnetic fields to childhood leukemia. *J Am Stat Assoc*. 2003;98(461):47-54. doi:10.1198/01621450338861905

86. Lash TL, Fink AK. Semi-automated sensitivity analysis to assess systematic errors in observational data. *Epidemiology*. 2003;14(4):451-458. doi:10.1097/01.EDE.0000071419.41011.cf

87. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *Int J Epidemiol*. 2005;34(6):1370-1376. doi:10.1093/ije/dyi184

88. Greenland S. Multiple-bias modelling for analysis of observational data (with discussion). *J R Stat Soc Ser A Stat Soc*. 2005;168(2):267-306. doi:10.1111/j.1467-985X.2004.00349.x

89. Lash TL, Schmidt M, Jensen AØ, Engebjerg MC. Methods to apply probabilistic bias analysis to summary estimates of association. *Pharmacoepidemiol Drug Saf*. 2010;19(6):638-644. doi:10.1002/pds.1938

90. Lash TL, Fox MP, Fink AK. *Applying Quantitative Bias Analysis to Epidemiologic Data*. Springer; 2009.

91. Orsini N, Bellocco R, Bottai M, Wolk A, Greenland S. A tool for deterministic and probabilistic sensitivity analysis of epidemiologic studies. *Stata J*. 2008;8(1):29-48. doi:10.1177/1536867x0800800103

92. Hunnicutt JN, Ulbricht CM, Chrysanthopoulou SA, Lapane KL. Probabilistic bias analysis in pharmacoepidemiology and comparative effectiveness research: A systematic review. *Pharmacoepidemiol Drug Saf*. 2016;25(12):1343-1353. doi:10.1002/pds.4076

93. Lash TL, Abrams B, Bodnar LM. Comparison of bias analysis strategies applied to a large data set. *Epidemiology*. 2014;25(4):576-582. doi:10.1097/EDE.0000000000000102

94. Ding P, VanderWeele TJ. Sensitivity analysis without assumptions. *Epidemiology*. 2016;27(3):368-377. doi:10.1097/EDE.0000000000000457

95. Smith LH, VanderWeele TJ. Bounding bias due to selection. *Epidemiology.* 2019;30(4):509-516. doi:10.1097/EDE.0000000000001032

96. VanderWeele TJ, Li Y. Simple sensitivity analysis for differential measurement error. *Am J Epidemiol.* 2019;188(10):1823-1829. doi:10.1093/aje/kwz133

97. Chin HB, Baird DD, McConnaughey DR, Weinberg CR, Wilcox AJ, Jukic AM. Long-term recall of pregnancy-related events. *Epidemiology.* 2017;28(4):575-579. doi:10.1097/EDE.0000000000000660

98. Greene N, Greenland S, Olsen JJ, et al. Estimating bias from loss to follow-up in the Danish national birth cohort. *Epidemiology.* 2011;22(6):1. doi:10.1097/EDE.0b013e31822939fd

99. Mumford SL, Yeung EH. Intergenerational effects–causation or confounding? *Fertil Steril.* 2018;110(1):52-53. doi:10.1016/j.fertnstert.2018.04.008

100. Omoni AO, Ntozini R, Evans C, et al. Child growth according to maternal and child HIV status in Zimbabwe. *Pediatr Infect Dis J.* 2017;36(9):869-876. doi:10.1097/INF.0000000000001574

101. Ross JA, Blair CK, Olshan AF, et al. Periconceptional vitamin use and leukemia risk in children with Down syndrome: A children's oncology group study. *Cancer.* 2005;104(2):405-410. doi:10.1002/cncr.21171

102. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med.* 2017;167(4):268-275. doi:10.7326/M16-2607

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

104. Maclure M, Schneeweiss S. Causation of bias: The episcope. *Epidemiology.* 2001;12(1):114-122. doi:10.1097/00001648-200101000-00019

105. Jurek AM, Maldonado G, Spector LG, Ross JA. Periconceptional maternal vitamin supplementation and childhood leukaemia: An uncertainty analysis. *J Epidemiol Community Health.* 2009;63(2):168-172. doi:10.1136/jech.2008.080226

106. Amitay EL, Keinan-Boker L. Breastfeeding and childhood leukemia incidence: A meta-analysis and systematic review. *JAMA Pediatr.* 2015;169(6):1-9. doi:10.1001/jamapediatrics.2015.1025

107. Mathur MB, Ding P, Riddell CA, VanderWeele TJ. Website and R package for computing E-values. *Epidemiology.* 2018;29(5):e45-e47. doi:10.1097/EDE.0000000000000864

108. Lash TL, Ahern TP. Bias analysis to guide new data collection. *Int J Biostat.* 2012;8(2):2. doi:10.2202/1557-4679.1345

109. Fox MP, Lash TL. Quantitative bias analysis for study and grant planning. *Ann Epidemiol.* 2020;43:32-36. doi:10.1016/j.annepidem.2020.01.013

110. Blum MR, Tan YJ, Ioannidis JPA. Use of E-values for addressing confounding in observational studies-an empirical assessment of the literature. *Int J Epidemiol.* 2020;49(5):1482-1494. doi:10.1093/ije/dyz261

111. Fox MP, Arah OA, Stuart EA. Commentary: The value of E-values and why they are not enough. *Int J Epidemiol.* 2020;49(5):1505-1506. doi:10.1093/ije/dyaa093

112. VanderWeele TJ, Mathur MB. Commentary: Developing best-practice guidelines for the reporting of E-values. *Int J Epidemiol.* 2020;49(5):1495-1497. doi:10.1093/ije/dyaa094

113. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol.* 1986;15(3):413-419. doi:10.1093/ije/15.3.413

114. VanderWeele TJ, Ding P, Mathur M. Technical considerations in the use of the E-value. *J Causal Inference.* 2019;7(2):1-11. doi:10.1515/jci-2018-0007

115. Ding P, Vanderweele TJ. Generalized Cornfield conditions for the risk difference. *Biometrika.* 2014;101(4):971-977. doi:10.1093/biomet/asu030

116. Frank KA. Impact of a confounding variable on a regression coefficient. *Sociol Methods Res.* 2000;29(2):147-194. doi:10.1177/0049124100029002001

117. Altonji JG, Elder TE, Taber CR. Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools. *J Polit Econ.* 2005;113(1):151-184. doi:10.1086/426036

118. Oster E. Unobservable selection and coefficient stability: Theory and evidence. *J Bus Econ Stat.* 2019;37(2):187-204. doi:10.1080/07350015.2016.1227711

119. Cinelli C, Hazlett C. Making sense of sensitivity: Extending omitted variable bias. *J R Stat Soc Ser B Stat Methodol.* 2019;82(1):39-67. doi:10.1111/rssb.12348

120. VanderWeele TJ. Optimal approximate conversions of odds ratios and hazard ratios to risk ratios. *Biometrics.* 2020;76(3):746-752. doi:10.1111/biom.13197

121. Phillips CV, LaPole LM. Quantifying errors without random sampling. *BMC Med Res Methodol.* 2003;3:1-10. doi:10.1186/1471-2288-3-9