2019

Essays In Causal Inference: Addressing Bias In Observational And Randomized Studies Through Analysis And Design

Raiden Berte Hasegawa
University of Pennsylvania, raiden.hasegawa@gmail.com

Essays In Causal Inference: Addressing Bias In Observational And Randomized Studies Through Analysis And Design

Abstract
In observational studies, identifying assumptions may fail, often quietly and without notice, leading to biased causal estimates. Although less of a concern in randomized trials where treatment is assigned at random, bias may still enter the equation through other means. This dissertation has three parts, each developing new methods to address a particular pattern or source of bias in the setting being studied. In the first part, we extend the conventional sensitivity analysis methods for observational studies to better address patterns of heterogeneous confounding in matched-pair designs. We illustrate our method with two sibling studies on the impact of schooling on earnings, where the presence of unmeasured, heterogeneous ability bias is of material concern. The second part develops a modified difference-in-difference design for comparative interrupted time series studies. The method permits partial identification of causal effects when the parallel trends assumption is violated by an interaction between group and history. The method is applied to a study of the repeal of Missouri's permit-to-purchase handgun law and its effect on firearm homicide rates. In the final part, we present a study design to identify vaccine efficacy in randomized control trials when there is no gold standard case definition. Our approach augments a two-arm randomized trial with natural variation of a genetic trait to produce a factorial experiment. The method is motivated by the inexact case definition of clinical malaria.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Statistics

First Advisor
Dylan S. Small

Keywords
Causal Inference, Observational Studies, Quasi-experimental Design, Randomized Trials, Sensitivity Analysis, Unobserved Confounding

Subject Categories
Statistics and Probability

ESSAYS IN CAUSAL INFERENCE: ADDRESSING BIAS IN OBSERVATIONAL AND RANDOMIZED STUDIES THROUGH ANALYSIS AND DESIGN

Raiden B. Hasegawa

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2019

Supervisor of Dissertation

Dylan S. Small, Class of 1965 Wharton Professor of Statistics

Graduate Group Chairperson

Catherine M. Schrand, Celia Z. Moh Professor of Accounting

Dissertation Committee

Paul R. Rosenbaum, Robert G. Putzel Professor of Statistics
Colin B. Fogarty, Assistant Professor, MIT Sloan School of Management
Bhaswar B. Bhattacharya, Assistant Professor of Statistics

ESSAYS IN CAUSAL INFERENCE: ADDRESSING BIAS IN OBSERVATIONAL AND RANDOMIZED STUDIES THROUGH ANALYSIS AND DESIGN

To my wonderful wife, Susanna, and my constant work companion, Poppy.

ACKNOWLEDGEMENT

To my advisor, Dylan, thank you for your guidance, tireless support, and friendship over these past five years. I will remember fondly our weekly meetings, your incredible ability to lay bare the components of a challenging problem in clear and simple terms, and your honest commitment to using statistics to make a real impact. I couldn't have asked for a more rewarding and supportive advising experience. Without you, this dissertation would not be possible.

To my committee, thank you. Paul, I could always count on you to ask questions that would make me think deeply about things I thought I understood. Bhaswar, it was a pleasure working closely with you as we taught the undergraduate horde! Colin, from fellow student to co-author to committee member, I'm grateful for your friendship and advice.

To my cohort, what a fun ride! Bikram, thanks for willingness to talk all things causal at all hours of the day. Linjun, thank you for your ever-friendly presence. Gemma and Justin, I'll miss our two-hour conference room lunches, but will take comfort in the lasting friendships that we've made. And also, Sameer, thank you for being a great friend, engaging collaborator, and expert mixologist.

To all the staff, faculty and students of the statistics department, thank you for fostering a collegial and caring environment. From day one, I've felt welcomed like a member of a big statistics family.

To my parents, Nancy and Doug, and my sister, Nika. I am lucky to call myself your son and brother. You knew me when I was a wild, noisy, and silly kid. You remind me that I can still be those things even when pursuing more "serious" endeavors. Nika, thank you for being my academic and life guide since day one. You led the way and all I had to do was try hard and follow, sometimes quite literally (Boola Boola!). Mom and Dad, you've been my earliest and most steadfast supporters in all I've done. I cannot say thank you enough

for the loving family you built for us. I love you all the big "I!"

To my parents-in-law, Ellen and Lars, thank you for welcoming me in to your family and supporting me throughout this process as a son.

To my furriest and most loyal supporter, Poppy, thank you for your companionship, goofy antics, and constant source of joy.

Finally, to the love of my life, my wife, Susanna. You were there with me every step of the way, making the tough times bearable and the successes all the more meaningful. That you were able to be my unflappable support system all while successfully completing an internal medicine residency astonishes me. Thank you for being my everything. I love you.

ABSTRACT

ESSAYS IN CAUSAL INFERENCE: ADDRESSING BIAS IN OBSERVATIONAL AND RANDOMIZED STUDIES THROUGH ANALYSIS AND DESIGN

Raiden B. Hasegawa

Dylan S. Small

In observational studies, identifying assumptions may fail, often quietly and without notice, leading to biased causal estimates. Although less of a concern in randomized trials where treatment is assigned at random, bias may still enter the equation through other means. This dissertation has three parts, each developing new methods to address a particular pattern or source of bias in the setting being studied. In the first part, we extend the conventional sensitivity analysis methods for observational studies to better address patterns of heterogeneous confounding in matched-pair designs. We illustrate our method with two sibling studies on the impact of schooling on earnings, where the presence of unmeasured, heterogeneous ability bias is of material concern. The second part develops a modified difference-in-difference design for comparative interrupted time series studies. The method permits partial identification of causal effects when the parallel trends assumption is violated by an interaction between group and history. The method is applied to a study of the repeal of Missouri's permit-to-purchase handgun law and its effect on firearm homicide rates. In the final part, we present a study design to identify vaccine efficacy in randomized control trials when there is no gold standard case definition. Our approach augments a two-arm randomized trial with natural variation of a genetic trait to produce a factorial experiment. The method is motivated by the inexact case definition of clinical malaria.

TABLE OF CONTENTS

ACKNOWLEDGEMENT iv

ABSTRACT vi

LIST OF TABLES x

LIST OF ILLUSTRATIONS xii

CHAPTER 1 : Introduction 1

CHAPTER 2 : Sensitivity Analysis for Matched Pair Analysis of Binary Data: From Worst Case to Average Case Analysis 4

2.1 Introduction 4

2.2 Notation and Review 10

2.3 From Worst Case to Average Case Sensitivity Analysis 17

2.4 The Effect of Cellphone Use on Motor-vehicle Collisions 19

2.5 Discussion 24

2.6 Appendix 25

CHAPTER 3 : Extended Sensitivity Analysis for Heterogeneous Unmeasured Confounding with an Application to Sibling Studies of Returns to Education 30

3.1 Introduction 30

3.2 Sensitivity analysis for paired studies 37

3.3 An extended sensitivity analysis 42

3.4 Implementation through quadratic programming 51

3.5 Simulations 55

3.6 Extended sensitivity analysis for returns to schooling 58

3.7 Concluding remarks 64
3.8 Appendix 66

CHAPTER 4 : Evaluating Missouri's Handgun Purchaser Law: A Bracketing Method for Addressing Concerns about History Interacting with Group . . 72
4.1 Comparative Interrupted Time Series Design and Potential Biases 72
4.2 Methods: Bracketing 75
4.3 Application: Effect of the Repeal of Missouri's Handgun Purchaser Licensing Law on Firearm Homicides 83
4.4 Conclusion and Discussion 89
4.5 Appendix 90

CHAPTER 5 : Estimating Malaria Vaccine Efficacy in the Absence of a Gold Standard Case Definition: Mendelian Factorial Design 101
5.1 Introduction 101
5.2 Mendelian Factorial Design: Parallels with Mendelian Randomization . . . 104
5.3 A Robust Framework for Estimating Vaccine Efficacy: Risk Ratios and Incidence Rate Ratios 107
5.4 Time-to-First Malaria Fever: Mendelian Factorial Design Under The Proportional Hazards Assumption 132
5.5 Discussion 139
5.6 Appendix 142

LIST OF TABLES

TABLE 1 : 2×2 contingency tables of cellphone use vs traffic collision incidence for different choices of control windows. 8

TABLE 2 : Sensitivity analysis for (marginal) $\alpha = 0.05$. 9

TABLE 3 : The four possible types of pairs in our case-crossover study. 14

TABLE 4 : Sensitivity analysis for 95% one-sided confidence intervals for attributable effects of the form $\{A : A > a^*\}$ 23

TABLE 5 : Sensitivity analysis for (marginal) $\alpha = 0.05$ and the expected lower bound on corresponding worst case calibrated bias. 27

TABLE 6 : Rejection probability of the true null hypothesis under biased setting. 57

TABLE 7 : Rejection probability of the false null hypothesis under unbiased setting and $H_1 : \tau = 0.5$. 58

TABLE 8 : 95% sensitivity intervals for $\log(\tau)$ in the AR study. 64

TABLE 9 : Rejection probability of the true null hypothesis under unbiased setting. 70

TABLE 10 : Rejection probability of the false null hypothesis under unbiased setting and $H_1 : \tau = 0.25$. 71

TABLE 11 : Age-adjusted firearm homicide rates for Missouri and Border States in pre-study, before repeal, and after repeal periods. 85

TABLE 12 : Difference-in-difference estimates of effect of repeal of Missouri's permit-to-purchase handgun licensing requirement on firearm homicide rates. 86

TABLE 13 : Difference-in-difference estimates of effect of repeal of Missouri's permit-to-purchase handgun licensing requirement on firearm homicide rates using after period of 2008-2013. 99

TABLE 14 : Parallel assumptions for Mendelian randomization and Mendelian factorial design. 105

TABLE 15 : Proportional absolute bias and root mean squared error (RMSE) of MFD and naive estimators using $N_{sim} = 5000$ simulations. 125

TABLE 16 : Coverage of two-sided 95% confidence interval and power against two-sided alternative at 5% significance level of MFD and naive estimators using $N_{sim} = 5000$ simulations. 126

LIST OF ILLUSTRATIONS

FIGURE 1 : Boxplots of differences in IQ scores between same-sex siblings where one attended college and the other did not. 33

FIGURE 2 : Histograms of between-sibling IQ disparities of same-sex sibling pairs and table of estimated increase in pairwise bias due to IQ disparities between siblings. 35

FIGURE 3 : Extended sensitivity curve from the AR study calibrated to the estimates of ability bias from the WLS study (cross). 61

FIGURE 4 : Histogram of $\boldsymbol{\pi}^*$ estimated for 171 same-sex, full-sibling pairs from the WLS study. 63

FIGURE 5 : Stylized plot of data from a comparative interrupted time series design. 73

FIGURE 6 : Age-adjusted firearm homicide rates in Missouri and states bordering Missouri (population-weighted averages), 1999-2016. 84

FIGURE 7 : Age-adjusted gun homicide rates in Missouri, lower control states, and upper control states, 1999-2016. 86

FIGURE 8 : Histograms of placebo "repeal" effects using upper and control states. 88

FIGURE 9 : Relative trends of homicide rates for Missouri, upper controls, and lower controls. 98

FIGURE 10 : Causal diagrams for MR (left) and MFD (right). 106

FIGURE 11 : 2×2 table for Mendelian factorial design. 116

FIGURE 12 : Distributions of simulated MFD estimator $\hat{\tau}$ and naive estimator $\hat{\tau}_0$ over several settings. 127

FIGURE 13 : Comparing densities and means of $\hat{\tau}_0$, $\hat{\tau}$, and $\hat{\tau}_{bnd}$ across 5000 simulations. 130

FIGURE 14 : Absolute proportional bias and RMSE for MFD, naive, and bounded estimators. 131
FIGURE 15 : Power against two-sided alternative and coverage for MFD, naive, and bounded estimators 132

CHAPTER 1

Introduction

The objective of many social science, epidemiology, and medical research studies is to identify and estimate causal relationships between treatments or exposures and outcomes of interest. In observational studies, the absence of physical randomization and the lack of the tightly controlled environment of a well planned experiment may lead critical observers to call causal conclusions into question. Although less of a concern in randomized trials where treatment is assigned at random, bias may still enter the equation through other means. For example, outcomes attributable to a disease of interest may be aliased with outcomes caused by other diseases when the symptoms associated with the disease are unspecific. Resulting case definitions are usually inexact and can lead to substantial bias even in otherwise well designed trials. Consequently, in both observational and randomized settings, anticipating and addressing plausible patterns of unmeasured confounding should be an objective of any research. In Chapters 2 through 5 we given four examples of how we address this objective through developments in both statistical *design* and *analysis*.

Chapters 2 and 3 approach the issue of bias in matched-pair studies through analysis, improving existing sensitivity analysis methods to more effectively address certain plausible patterns of bias.

In Chapter 2, we introduce a sensitivity analysis framework that allows for the investigator to interpret the sensitivity parameter as a bound on the average bias present in a matched-pair study with binary outcomes (Hasegawa and Small, 2017). The new interpretation resolves difficulties of the standard sensitivity analysis that bounds the maximal bias to which pairs are subject, when the pattern of bias is presumed to be heterogeneous (Rosenbaum, 1987). Specifically, when some pairs may suffer from arbitrarily large biases, but on the average the study is more moderately biased, the average case sensitivity analysis will be preferable to the standard approach. We motivate the method with a study of the effects

of talking on a mobile phone on the incidence of car accidents (Tibshirani and Redelmeier, 1997).

In Chapter 3, we extend this framework using modern convex optimization tools to allow for continuous outcomes and a simultaneous bound on both the maximal and average (or typical) bias in a matched pair study. We call this the *extended sensitivity analysis* framework (Fogarty and Hasegawa, 2019). In addition to bounding sample-level bias, extended sensitivity analysis lets the investigator place bounds on the typical bias present in a superpopulation from which the paired sample was drawn. This allows for calibration of a sensitivity analysis in one study to information on confounding from another study whose sample was generated from the same superpopulation as the first. We apply these new methods to two sibling studies on the effects of education on future earnings. We calibrate the extended sensitivity of one study where IQ data was not collected to an estimate of bias introduced by differences in ability between siblings from the second study where IQ data was collected. Empirically, the example suggests that ability bias is heterogeneous across sibling pairs; the bias is typically modest but there is a small proportion of sibling pairs where the differences in ability are quite large. We demonstrate that the extended sensitivity analysis is better suited than the standard sensitivity analysis in such settings.

Through design, Chapter 3 addresses bias in comparative interrupted times series when the parallel trends assumption of the standard difference-in-difference design fails because of an interaction between history and the groups under comparison (Hasegawa et al., 2019). We develop a difference-in-difference based design that allows for partial identification of causal effects when the parallel trends assumption fails. We re-analyze a study of the repeal of Missouri's permit-to-purchase law and its effect on firearm homicide rates (Webster et al., 2014) using our prosed method. The repeal occurred concurrently with the Great Recession and there is concern that the firearm homicide trends in Missouri and the control states may have been differentially affected by the onset of recession. Our method provides partial identification of the repeal effect under mild assumptions about how the recession interacted

with the different groups.

Finally, in Chapter 5, we give another example of how improved design can mitigate concerns of bias, this time in a randomized control trial to assess the efficacy of a malaria vaccine. Symptoms of clinical malaria have significant overlap with the symptoms of other common childhood illnesses. Furthermore, children in endemic areas are able to tolerate varying levels of parasitemia without symptoms. Together, these facts make distinguishing between malaria-attributable symptoms and non-malaria symptoms very challenging. Inexact case definitions currently in use can substantially bias estimates of vaccine efficacy. In this chapter, we leverage genetic traits that are protective against malaria but not other childhood illnesses to identify vaccine efficacy in a randomized control trial. The sickle cell trait is one such genetic variant that confers protection specifically against clinical malaria. The method, which we call *mendelian factorial design*, is inspired by mendelian randomization studies that use genetic variants as instrumental variables to estimate causal effects of non-randomized exposures. Under realistic assumption, this new study design allows for identification of vaccine efficacy.

CHAPTER 2

Sensitivity Analysis for Matched Pair Analysis of Binary Data: From Worst Case to Average Case Analysis

Abstract

In matched observational studies where treatment assignment is not randomized, sensitivity analysis helps investigators determine how sensitive their estimated treatment effect is to some unmeasured confounder. The standard approach calibrates the sensitivity analysis according to the worst case bias in a pair. This approach will result in a conservative sensitivity analysis if the worst case bias does not hold in every pair. In this paper, we show that for binary data, the standard approach can be calibrated in terms of the average bias in a pair rather than worst case bias. When the worst case bias and average bias differ, the average bias interpretation results in a less conservative sensitivity analysis and more power. In many studies, the average case calibration may also carry a more natural interpretation than the worst case calibration and may also allow researchers to incorporate additional data to establish an empirical basis with which to calibrate a sensitivity analysis. We illustrate this with a study of the effects of cellphone use on the incidence of automobile accidents. Finally, we extend the average case calibration to the sensitivity analysis of confidence intervals for attributable effects.

2.1. Introduction

2.1.1. Sensitivity analysis as causal evidence

In matched-pair observational studies, causal conclusions based on usual inferential methods (e.g., McNemar's test for binary data) rest on the assumption that matching on observed covariates has the same effect as randomization (i.e., that there are no unmeasured confounders). In other words, it is assumed that there are no unobserved covariates relevant to both treatment assignment and outcome. A sensitivity analysis assesses the sensitivity of results to violations of this assumption. Cornfield et al. (1959) introduced a model for sensitivity analysis that was a major conceptual advance in the field of observational studies. A

modern approach to sensitivity analysis is introduced in Rosenbaum (1987); Rosenbaum's approach builds on Cornfield's model (Cornfield et al. (1959)) but incorporates uncertainty due to sampling variance. There are other contemporary sensitivity analysis models, see for example McCandless et al. (2007) for a Bayesian approach, but we restrict our focus to Rosenbaum's approach. Rosenbaum's sensitivity analysis yields an upper limit on the magnitude of bias to which the result of the researcher's test of no treatment effect is insensitive for a given significance level α. More specifically, Rosenbaum (1987) derives bounds on the p-value of this test given an upper bound, Γ, on the odds ratio of treatment assignment for a pair of subjects matched on observed covariates. Γ can be thought of as a measure of "worst case" bias in the sense that treatment assignment probabilities in matched pairs are allowed to vary arbitrarily as long as the odds ratio of treatment assignment for a pair of subjects is no greater than Γ. The largest Γ for which the p-value is less than 0.05 is denoted by Γ_{sens}. We will use Γ_{truth} to distinguish the true unknown worst case bias. Γ_{sens} is interpreted in Rosenbaum's sensitivity analysis as the largest value of the worst case bias across matched pairs that does not invalidate the finding of evidence for a treatment effect. We refer to this as a *worst case calibrated* sensitivity analysis. A classic example of this type of analysis is given in Chapter 4 of Rosenbaum (2002c). Applying the worst case sensitivity analysis to a study of the effects of heavy smoking on lung cancer mortality (Hammond (1964)), Rosenbaum finds that $\Gamma_{sens} \approx 6$ and interprets this result cogently:

> To attribute the higher rate of death from lung cancer to an unobserved covariate rather than to an effect of smoking, that unobserved covariate would need to produce a sixfold increase in the odds of smoking, and it would need to be a near perfect predictor of lung cancer.

A brief, more formal review of Rosenbaum's sensitivity analysis framework is in Section 2.2.2.

The worst case calibrated sensitivity analysis raises several potential questions. If we are convinced that there is no pair in Hammond's smoking study such that one unit is more than

six times as likely to smoke as the other (i.e., $\Gamma_{truth} \leq \Gamma_{sens}$), then we would conclude that our study provides convincing evidence that heavy smoking increases the rate of lung cancer mortality. However, what if, on average, unmeasured confounders do not alter the odds of smoking greatly but there are some subjects for whom the unmeasured confounders make them almost certain to smoke, e.g., a subject who experiences huge peer pressure to smoke. If such a subject ends up in our sample of matched pairs, and we condition on matched pairs in which only one unit receives treatment, a standard practice when conducting matched pair randomization tests, then the odds ratio of treatment assignment in the matched pair containing that subject, and consequently Γ_{truth}, will be infinite. In such a case, since Γ_{sens} is generally finite, we'd expect it to be smaller than Γ_{truth}. Now, suppose that there are such pairs in the Hammond study but that for most pairs the odds ratio of smoking between the units is much smaller than six. Using the worst case calibrated sensitivity analysis, we would conclude that the study is sensitive to bias. Is there potentially some natural quantification of average bias over the sample of matched pairs, say, Γ'_{truth}, that isn't infinite and perhaps is smaller than six? And if we calibrate our sensitivity analysis to this measure of bias rather than the worst case measure, will the sensitivity analysis be valid in the sense that the inference is conservative at level α for any $\Gamma \geq \Gamma'_{truth}$? If it is valid, are there other advantages to using the *average case calibrated* sensitivity analysis over the worst case calibrated sensitivity analysis? In what follows, we attempt to answer these motivating questions in the context of a matched pair analysis of the association between cellphone use and car accidents.

2.1.2. Outline

In this paper we demonstrate that interpreting sensitivity analysis results in terms of average case rather than worst case hidden bias is both valid and conceptually more natural in many common scenarios. To illustrate our claim that the average case analysis is more natural we will perform a causal analysis of a study by Tibshirani and Redelmeier (1997) that asks if there is an association between cellphone use and motor-vehicle collisions. The study is

described in the following section. In section 2.2 we review the model for sensitivity analysis of tests of no treatment effect and sensitivity intervals for attributable effects for binary data. In section 2.3 we discuss the theory behind the validity of average case sensitivity analysis. Finally, the Tibshirani and Redelmeier (1997) study is examined in this new light in section 2.4. In particular, we see how the average case sensitivity analysis makes it possible to use additional information from the problem to empirically calibrate our sensitivity analysis in Section 2.4.1 and we extend the average case sensitivity analysis to the study of sensitivity intervals for attributable effects in Section 2.4.3.

2.1.3. Motivating Example: Effects of cellphone use on the incidence of motor-vehicle collisions

Tibshirani and Redelmeier (1997) conducted a case-crossover study of the effects of cellphone use on the incidence of car collisions. In a case-crossover study each subject acts as her own control which has the benefit of controlling for potential confounders that are time-invariant, even if they are unobserved. Data collection took place at a collision reporting center in Toronto between July 1, 1994 and August 31, 1995 during weekday peak hours (10 AM to 6 PM). Consenting drivers who reported having been in a collision with substantial property damage and who owned a cellphone were included in the study. Drivers involved in collisions that involved injury, criminal activity, or transport of dangerous goods were excluded. The resulting study population included 699 individuals who gave permission to review their cellphone records and filled out a brief questionnaire about their personal characteristics and the features of the collision. The matched pair analysis compared cellphone usage in the 10-minute hazard window prior to the crash with a 10-minute control window on a chosen day prior to the crash. We will denote the time of the crash as t and the hazard window as $t-10$ to $t-1$ minutes. The authors examined several different control windows:

1. *Previous day*: time $t-10$ to $t-1$ minutes on the previous day.

2. *Previous weekday/weekend*: time $t-10$ to $t-1$ minutes on the previous weekday if the

crash took place on a weekday and similarly if the crash took place on a weekend.

3. *One week prior*: time $t-10$ to $t-1$ minutes one week prior to the collision.

4. *Busiest cellphone day of previous three days*: time $t-10$ to $t-1$ minutes on the one day among the prior three to the collision with the most cellphone calls.

For each choice of control window, Tibshirani and Redelmeier (1997) found that there was a significant positive association between cellphone usage and traffic collision incidence. The 2 x 2 contingency tables shown in Table 1 summarize the data using the four different control windows.

		Control	
		On phone	Not on phone
		Previous Weekday/end	
Hazard	On phone	12	158
	Not on phone	23	506
		One Week Prior	
Hazard	On phone	6	164
	Not on phone	21	508
		Previous Driving Day	
Hazard	On phone	18	119
	Not on phone	20	171
		Most Active Cellphone Day	
Hazard	On phone	17	135
	Not on phone	43	504

Table 1: **One Week Prior**: results for one week prior control window versus hazard window;**Previous Weekday/end**: results for previous weekday/weekend control window versus hazard window; **Previous Driving Day**: results for previous driving day control window versus hazard window;**Most Active Cellphone Day**: results for most active cellphone day in previous 3 days control window versus hazard window.

2.1.4. Sensitivity of results to hidden bias

As this was an observational study, the associations cannot be assumed to be causal. We would like to quantify how large a hidden bias would have to be to explain the observed association between cellphone use and car accidents without it being causal. A sensitivity

analysis seems appropriate and is a straightforward exercise (see Chapter 4, Rosenbaum (2002c) for example). Table 2 shows the results of a standard worst case sensitivity analysis for each control window. Here, Γ_{sens} is the largest value of Γ such that the result are still significant at the $\alpha = 0.05$ level. In our analysis of the case-crossover study from Tibshirani and Redelmeier (1997) we condition on subjects who were on a cellphone in exactly one of the control and hazard windows (i.e., discordant case-crossover pairs). Thus, the odds ratio of treatment assignment for the two windows observed for any case-crossover subject can be viewed as the conditional odds that treatment occurs in a particular window. Hence, we can interpret Γ as the maximum (and $1/\Gamma$ as the minimum) over all study subjects of the odds that a driver is using a cellphone during the hazard window and not during the control window.

Control Window	Γ_{sens}
previous weekday/weekend	4.92
one week prior	5.53
previous driving day	4.15
most active cellphone day	2.40

Table 2: Sensitivity analysis for (marginal) $\alpha = 0.05$.

The sensitivity analysis suggests that the most active cellphone day control window was the most conservative analysis. This is unsurprising since we would expect that the treatment assignment (cellphone use) would be biased toward the control window on a day when you used a cellphone relatively often. We can interpret these results as follows: *the observed ostensible effect is insensitive to hidden bias that increases the odds that a driver was on a cellphone in the hazard window and not the control window on the most active cellphone day by at most a factor of 2.4.* In many observational studies this type of statement is very useful. However, it may be plausible that some study participants are exposed to infinite (or at least very large) hidden bias. For example, this happens if a subject was not driving during the control window and (almost) always uses her landline rather than her cellphone when she is not driving. When we condition on case-crossover pairs where the

treatment is received in exactly one of the windows – a standard practice when conducting a matched pair randomization test – such a driver is always on a cellphone during the hazard window. When this happens, the observed ostensible effect is (almost) always sensitive to hidden bias, no matter how strong the observed association. Implicitly, in the worst case sensitivity analysis, the investigator is supremely skeptical; she assumes that it could be that all study participants suffer from the worst case hidden bias which, when it is possible that some study participant suffers from unbounded hidden bias, renders sensitivity analysis under the standard worst case interpretation uninformative. Yet in many studies where unbounded hidden bias in some matched pairs is plausible, as in our motivating example, we still want to examine the sensitivity of our results to potential hidden bias. If we could perform a valid, average case calibrated sensitivity analysis then we could (1) make sensitivity analysis informative even in the presence of pairs subject to unbounded hidden bias and (2) make the interpretation of sensitivity analysis results far less conservative. It turns out that there is a measure of the sample average bias that is generally finite in the presence of pairs subject to unbounded bias for data with binary treatment and outcome. Moreover, the sensitivity analysis calibrated to this measure of average bias is valid when using McNemar's statistic to test the null hypothesis of no treatment effect against the alternative of a positive treatment effect (i.e., that talking on a cellphone while driving increases the rate of automobile accidents).

2.2. Notation and Review

2.2.1. Notation

Our study sample consists of S matched pairs where each pair $s = 1, 2, \ldots, S$ is matched on a set of observed relevant covariates $\mathbf{x}_{s1} = \mathbf{x}_{s2} = \mathbf{x}_s$. Units in each pair are indexed by $i = 1, 2$. We let Z_{si} and R_{si} denote the treatment assignment and outcome, respectively, of the i-th unit of the s-th pair. The potential outcomes under treatment and control are denoted as r_{Tsi} and r_{Csi}, respectively. Hence, we can write $R_{si} = Z_{si}r_{Tsi} + (1 - Z_{si})r_{Csi}$. Under Fisher's sharp null hypothesis of no treatment effect, i.e., $r_{Tsi} = r_{Csi}$ for all i, we have that

$R_{si} = r_{Csi}$. Hereafter, we will work under the null hypothesis and under the assumption that each pair was matched on some set of observed covariates $\mathbf{x}_s$. Additionally, we assume that there is some unobserved covariate U_{si} that is associated with both treatment assignment and outcome and let u_{si} be the realization of U_{si} for the i-th unit of the s-th pair. Within pair differences in treatment and outcome will be denoted as $V_s = Z_{s1} - Z_{s2}$ and $y_s = r_{Cs1} - r_{Cs2}$. It will be convenient to define the following vector quantities: $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{S2})^T$, $\mathbf{r} = (r_{C11}, r_{C12}, \ldots, r_{CS2})^T$,$\mathbf{U} = (U_{11}, U_{12}, \ldots, U_{S2})^T$, and $\mathbf{A} = (|y_1|, |y_2|, \ldots, |y_S|)^T$.

To be very clear about the information on which we are conditioning we will define some important information sets. Let $\mathcal{F} = \{(\mathbf{x}_s, u_{si}, r_{Csi}, r_{Tsi}) : \ s = 1, 2 \ldots, S, \, i = 1, 2\}$ be the set of *fixed* observed and unobserved covariates for all units. Let $\mathcal{Z} = \{\mathbf{Z} : \ |V_s| = 1, \, s = 1, \ldots, S\}$ be the set of matched pairs such that only one unit receives treatment. We assume that $\mathbf{R}$ is binary and we define $\mathcal{A}_\mathbf{1} = \{\mathbf{A} : \ |y_s| = 1, \, s = 1, \ldots, S\}$. So $\mathcal{Z} \cap \mathcal{A}_\mathbf{1}$ is the set of discordant matched pairs. In the analysis that follows, we will condition on $\mathcal{F}$, $\mathcal{Z} \cap \mathcal{A}_\mathbf{1}$.

2.2.2. Review: sensitivity analysis for binary data

Under the assumption that all variables that confound treatment assignment are observed,

$$Z_{si} \perp\!\!\!\perp (r_{Csi}, r_{Tsi}) \,|\, \mathbf{X}_s \quad \text{(Ignorability)}$$

our matched observational study should closely resemble a randomized study and thus $\mathbb{P}(\mathbf{Z} = \mathbf{z} | \, \mathcal{F}, \, \mathcal{Z} \cap \mathcal{A}_\mathbf{1}) = 1/2^S$ for $\mathbf{z} \in \mathcal{Z}$. In practice, this assumption is rarely valid and the probability of treatment assignment depends materially on the unobserved covariates $\mathbf{U}$. A second assumption made in the causal framework introduced in Rosenbaum and Rubin (1983) is the *Positivity* assumption – $0 < \mathbb{P}(Z_{si} = 1 | \, \mathbf{X}_s) < 1$ for all $s = 1, 2, \ldots, S$ and $i = 1, 2$ – which says that all units have a chance of receiving treatment. In our case-crossover study, however, this may not be an appropriate assumption. We introduce an example of how our case-crossover study might violate the positivity assumption in Section 2.4.1 and how our average case sensitivity analysis framework is able to handle violations

of positivity.

When both Z and r are binary it is common to use McNemar's statistic to test for treatment effect:

Definition 1. *For a matched pair study with binary treatment and outcome we define* ***McNemar's statistic*** *to be*

$$T(\mathbf{Z}, \mathbf{r}) = \sum_{s=1}^{S} \mathbb{1}\{V_s Y_s = 1\}\,. \tag{2.1}$$

Under the null distribution of no treatment effect $T(\mathbf{Z}, \mathbf{r})$ follows a Poisson-Binomial distribution with probabilities $\{p_1, p_2, \ldots, p_S\}$ where $p_s = \mathbb{P}((Z_{s1} - Z_{s2})(r_{s1} - r_{s2}) = 1)$ is the probability that the unit with positive outcome, i.e., $r = 1$, receives treatment in pair s. If we consider only discordant pairs and we assume, without loss of generality, that the first unit in each pair is the unit with positive outcome we may write

$$p_s = \mathbb{P}(Z_{s1} = 1 | \mathcal{F},\, \mathcal{Z} \cap \mathcal{A}_{\mathbf{1}})\ . \tag{2.2}$$

Recall that the Poisson-Binomial distribution is the sum of independent, not necessarily identical Bernoulli trials. If $\mathbf{X}_s$ contains the complete set of relevant covariates then p_s equals $1/2$ for all pairs and we can conduct inference using $\mathrm{B}(1/2, S)$ as our null distribution, effectively treating our data as being the outcome of a randomized study. As we mentioned earlier in this section, if there is some unobserved characteristic U that is relevant to treatment assignment and outcome then $\{p_1, \ldots, p_S\}$ are unknown and consequently the exact null distribution is no longer available to the investigator. When this is the case, a sensitivity analysis like the one conducted informally in Section 2.1.4 can be used to determine how sensitive the investigator's conclusions are to departures from the ideal randomized design. Following Chapter 4 of Rosenbaum (2002c) we can formalize the notion of a sensitivity analysis introduced in Sections 2.1.1 and 2.1.4 with a simple sensitivity model

where

$$\frac{1}{1+\Gamma} \leq \mathbb{P}(Z_{s1} = 1 | \mathcal{F},\, \mathcal{Z} \cap \mathcal{A}_{\mathbf{1}}) \leq \frac{\Gamma}{1+\Gamma} \tag{2.3}$$

for all $s = 1, \ldots, S$ and where $\Gamma \geq 1$ is the sensitivity parameter that bounds the extent of departure from a randomized study. Proposition 12 in Chapter 4 of Rosenbaum (2002c) states that (2.3) is equivalent to the existence of the following model

$$\log\left(\frac{p_s}{1-p_s}\right) = \gamma\left(u_{s1} - u_{s2}\right)\ ,\ s = 1, \ldots, S \tag{2.4}$$

where $\exp(\gamma) = \Gamma$, $\gamma \geq 0$, and $u_{si} \in [0,1]$ for $s = 1, \ldots, S$ and $i = 1, 2$. The restriction of the unobserved confounder to the unit interval in this equivalent representation preserves the non-technical interpretation of Γ used in section 2.1.4 as a bound on the odds that the driver was talking on a cellphone in the hazard window. Henceforth, we assume that U_{si} and its realization u_{si} belongs to the unit interval for $s = 1, \ldots, S$ and $i = 1, 2$. However, the distribution of U_{si} on the unit interval may be arbitrary.

Under this sensitivity model, if we let T^+ be binomial with success probability $\Gamma/(1+\Gamma)$ and T^- be binomial with success probability $1/(1+\Gamma)$ it follows from Theorem 2 of Rosenbaum (1987) that

$$\mathbb{P}\big(T^- \geq k\big) \leq \mathbb{P}(T \geq k | \mathcal{F},\, \mathcal{Z} \cap \mathcal{A}_{\mathbf{1}}) \leq \mathbb{P}\big(T^+ \geq k\big) \tag{2.5}$$

for all $k = 1, \ldots, S$. This inequality is tight in the sense that it holds for any realization $\mathbf{u}$ of $\mathbf{U}$. For conducting a hypothesis test, the stochastic ordering in (2.5) gives us bounds on the p-value of our test for a given magnitude of bias Γ. If $\Gamma \geq \Gamma_{truth}$, then T^+ yields a valid, albeit conservative, reference distribution for testing the null hypothesis of no treatment effect against the alternative of a positive treatment effect.

2.2.3. Attributable effects for binary outcomes: hypothesis tests and confidence intervals

Attributable effects are a way to measure the magnitude of a treatment effect on a binary outcome. The number of attributable effects is the number of positive outcomes among

treated subjects that would not have occurred if the subject was not exposed to treatment. In this section, we review Rosenbaum (2002a)'s procedure to construct one-sided confidence statements about attributable effects in the context of the cellphone case-crossover study.

Let $\widetilde{S}$ be the number of *all* pairs in the study, discordant or not, and let the first S be the discordant pairs. If we assume that $r_{Tsi} \geq r_{Csi}$, that talking on a cellphone cannot prevent an accident, then we can write the attributed effect as

$$A = \sum_{s=1}^{\widetilde{S}} \sum_{i=1}^{2} Z_{si}(r_{Tsi} - r_{Csi}) = \sum_{s=1}^{\widetilde{S}} Z_{s1}(r_{Ts1} - r_{Cs1}) \tag{2.6}$$

where the first unit of s-th pair is the observation from the hazard window. Why does the second equality hold? If the subject was talking on a cellphone in the control window, that is $Z_{s2} = 1$, then we observe $r_{Ts2} = 0$ which by our assumption that talking on a cellphone cannot prevent an accident implies that $r_{Cs2} = 0$. So attributable effects can only occur among discordant pairs where the subject was talking on a cellphone in the hazard window or concordant pairs where the subject was talking on a cellphone in both windows. The following table characterizes the four types of possible pairs in our case-crossover study,

	Z_{s1}	Z_{s2}	R_{s1}	R_{s2}	r_{Ts1}	r_{Cs1}
$D(+,-)$	1	0	1	0	1	-
$D(-,+)$	0	1	1	0	1	1
$C(-,-)$	0	0	1	0	1	1
$C(+,+)$	1	1	1	0	1	-

Table 3: The four possible types of pairs in our case-crossover study. D and C indicate discordant and concordant pairs, respectively, and the $+$ and $-$ indicate if a unit in the pair was treated or not, respectively.

D and C indicate discordant and concordant pairs, respectively. $D(+,-)$ is the set of discordant pairs where the subject was on a cellphone in the hazard window, $D(-,+)$ is the set of discordant pairs where the subject was on a cellphone in the control window, $C(+,+)$ is the set of concordant pairs where the subject was on a cellphone in both hazard and control windows, and $C(-,-)$ is the set of concordant pairs where the subject was not

on a cellphone in either window. If there are no attributable effects then we know that $r_{Cs1} = 1$ in $D(+,-)$ and $C(+,+)$ and we have that $R_{s1} = r_{Cs1}$ for all pairs s, concordant or discordant. We can write the probability that the subject was talking on a cellphone at the time of accident for each type of pair as (1) $\mathbb{P}(Z_{s1}R_{s1} = 1|D(+,-) \cup D(-,+)) = p_s$, where p_s here is equivalent to the p_s defined in Section 2.2.2 when there are no attributable effects; (2) $\mathbb{P}(Z_{s1}R_{s1} = 1|C(-,-)) = 0$; and (3) $\mathbb{P}(Z_{s1}R_{s1} = 1|C(+,+)) = 1$. Now let $c^+ = |C(+,+)|$ denote the cardinality of the set of concordant pairs where the subject was on a cellphone in both windows and let $s = S+1, \ldots, S+c^+$ be the pairs belonging to $C(+,+)$. Then if $A = 0$ we can define the standardized deviate for McNemar's statistic T as

$$\begin{aligned} \widetilde{T} &= \frac{\sum_{s=1}^{S} Z_{s1} r_{Cs1} - \sum_{s=1}^{S} p_s}{\left\{\sum_{s=1}^{S} p_s(1-p_s)\right\}^{1/2}} \\ &= \frac{\sum_{s=1}^{S+c^+} Z_{s1}R_{s1} - \left(\sum_{s=1}^{S} p_s + c^+\right)}{\left\{\sum_{s=1}^{S} p_s(1-p_s)\right\}^{1/2}}. \end{aligned} \tag{2.7}$$

$\widetilde{T}$ defines a normal reference distribution for $\sum_{s=1}^{S} Z_{s1} r_{Cs1}$ that we can use to conduct approximate inference. If $A = a > 0$, then $Z_{\widetilde{s}1}R_{\widetilde{s}1} = Z_{\widetilde{s}1}r_{T\widetilde{s}1} = Z_{\widetilde{s}1}(r_{C\widetilde{s}1} + 1)$ for pair $\widetilde{s}$ belonging to the set of a pairs with attributable accidents and the second equality above does not hold. When this equality fails to hold, the standard normal deviate $\widetilde{T}$ cannot be computed from the observed data conditional on $\mathcal{F}$. How then can we adjust $\widetilde{T}$ for attributable accidents so that it can be computed from the observed data? Because we've assumed talking on a cellphone cannot prevent an accident, we only need to consider two cases. If pair $\widetilde{s}$ belongs to $D(+,-)$ then we subtract $Z_{\widetilde{s}1}(r_{T\widetilde{s}1} - r_{C\widetilde{s}1}) = 1$ from $\sum_{s=1}^{\widetilde{S}} Z_{s1}R_{s1}$, $p_{\widetilde{s}}$ from the expectation, and $p_{\widetilde{s}}(1-p_{\widetilde{s}})$ from the variance term. If $\widetilde{s}$ belongs to $C(+,+)$ we again subtract 1 from $\sum_{s=1}^{\widetilde{S}} Z_{s1}R_{s1}$ and subtract 1 from the $|C(+,+)|$ in the expectation while leaving the variance term unchanged.

Let $\boldsymbol{\delta} = (\delta_{11}, \delta_{12}, \ldots, \delta_{\widetilde{S}1}, \delta_{\widetilde{S}2})^T$ be defined as $\delta_{sj} = r_{Tsj} - r_{Csj}$. We say that $\boldsymbol{\delta}$ is *compatible*

if $\delta_{sj} = 0$ whenever $Z_{sj} = 1$ and $R_{sj} = 0$ or $Z_{sj} = 0$ and $R_{sj} = 1$. Under this definition, we can express the number of attributable effects as $A = \mathbf{Z}^T\boldsymbol{\delta}$. For a compatible $\boldsymbol{\delta}$ such that $\mathbf{Z}^T\boldsymbol{\delta} = a$ we denote $\widetilde{T}_{-\boldsymbol{\delta}}$ to be $\widetilde{T}$ adjusted for the a attributable effects. $\widetilde{T}_{-\boldsymbol{\delta}}$ defines a new reference distribution for $\sum_{s=1}^{S} Z_{s1}r_{Cs1}$ under the null hypothesis that potential accidents indicated by $\boldsymbol{\delta}$ are attributable to talking on a cellphone while driving. We can write $\widetilde{T}_{-\boldsymbol{\delta}}$ as

$$\widetilde{T}_{-\boldsymbol{\delta}} = \frac{\sum_{s=1}^{S+c^+} Z_{s1}R_{s1}(1-\delta_{s1}) - \left(\sum_{s=1}^{S}(1-\delta_{s1})p_s + \sum_{s=S+1}^{S+c^+}(1-\delta_{s1})\right)}{\left\{\sum_{s=1}^{S}(1-\delta_{s1})p_s(1-p_s)\right\}^{1/2}}. \tag{2.8}$$

Using the notion of asymptotic separability (Gastwirth et al. (2000)), Rosenbaum (2002a) show that choosing a compatible $\boldsymbol{\delta}^* \equiv \boldsymbol{\delta}^*(a)$ with $\mathbf{Z}^T\boldsymbol{\delta}^*(a) = a$ that maximizes the expectation, and when there are ties to maximize the variance term, yields a reference distribution that, asymptotically, has the largest upper tail area among compatible $\boldsymbol{\delta}(a)$. Thus, we can use $T_{-\boldsymbol{\delta}^*}$ to test the plausibility that there are at most a attributable effects. Since A is a random variable we refrain from calling this a hypothesis test, a term usually reserved for unknown parameters. From equation (2.8) we see that $\boldsymbol{\delta}^*(a)$ includes the a pairs in $D(+,-)$ with the smallest values of p_s.

It is possible to invert the one-sided "plausibility tests" introduced above using $T_{-\boldsymbol{\delta}^*}$ that we just introduced in order to construct a confidence interval for attributable effects of the form $\{A : A > a\}$. It turns out that if it is plausible that there are a attributable effects then it is also plausible that there are $a+1$ attributable effects (Rosenbaum (2002c)). This monotonicity property leads to a very simple procedure to construct a one-sided confidence interval in the absence of hidden bias. First, if $p_s = 1/2$ for all $s = 1, 2, \ldots, \widetilde{S}$ then for any $a \geq 0$ we can compute $\widetilde{T}_{-\boldsymbol{\delta}^*} = \{T - a - (S-a)/2\}/\{(S-a)^{1/2}/2\}$.

Next, starting with $a = 0$ we check if $\widetilde{T}_{-\boldsymbol{\delta}^*} < \Phi^{-1}(1-\alpha)$, incrementing a by one if it isn't and stopping if it is. Finally, let a^* be equal to one less the value of a at which we terminate the procedure. Using the monotonicity result above we have that $\{A : A > a^*\}$ is a one-sided $100 \times (1-\alpha)\%$ confidence interval.

If we bound the worst case calibrated bias above by Γ then we can construct a one-sided $100 \times (1-\alpha)\%$ confidence interval following the same procedure but instead using $\widetilde{T}_{-\boldsymbol{\delta}^*,\Gamma} = \{T - a - (S-a)p_\gamma\}/\{(S-a)p_\gamma(1-p_\gamma)\}^{1/2}$ as our standard deviate where $p_\gamma = \Gamma/(1+\Gamma)$. The resulting one-sided $100 \times (1-\alpha)\%$ confidence interval is referred to as a *sensitivity interval* (See Chapter 4, Rosenbaum (2002c)). For a detailed illustration of these procedures we refer the reader to Sections 3-6 of Rosenbaum (2002a).

2.3. From Worst Case to Average Case Sensitivity Analysis

2.3.1. Valid average case analysis: binary outcome

An investigator conducting a sensitivity analysis tries to determine a test statistic whose null distribution is known conditional on the presence of hypothetical bias Γ. Since the distribution of U_{si} is unknown, traditionally, the investigator assumes the worst. That is, the null distribution is constructed assuming that in each pair $u_{s1} = 1$ and $u_{s2} = 0$. As noted in Section 2.2.2, T^+ yields a valid reference distribution for testing the null of no-treatment effect when $\Gamma \geq \Gamma_{truth}$. However, such a test is inherently conservative because it is designed to be valid for any realization of $\mathbf{U}$ since $\mathbf{U}$ and thus since $\mathbf{p} = (p_1, \ldots, p_S)^T$ are generally unknown. This is why we resort to a sensitivity analysis where we allow p_s to vary arbitrarily as long as $p_s/(1-p_s) \leq \Gamma$. In Section 2.1.1 we asked whether there was some natural quantification of average bias to which we could calibrate our sensitivity analysis which would lead to a less conservative analysis than the worst case calibration. One such quantification is $\Gamma'_{truth} = \overline{\mathbf{p}}/(1-\overline{\mathbf{p}})$ where $\overline{\mathbf{p}}$ is the sample average of p_s. In what follows, we show that if we calibrate our sensitivity analysis to Γ'_{truth} it will be valid and less conservative than the worst case calibration. To prove this, we show that $T' \sim \mathrm{B}(\Gamma'_{truth}/(1+\Gamma'_{truth}), S)$ yields a valid reference distribution for testing the null of no treatment effect against the alternative of a positive treatment effect. In Theorem (2) below, we prove that the upper tail probability for McNemar's statistic T is bounded above by the upper tail probability for T'.

Theorem 2. *Set* $\overline{\mathbf{p}} = \left(\sum_{s=1}^{S} p_s\right)/S$ *and* $\Gamma'_{truth} = \overline{\mathbf{p}}/(1-\overline{\mathbf{p}})$ *and let* $V_s \overset{iid}{\sim} \mathrm{Bern}(\Gamma'_{truth}/(1+$

Γ'_{truth})) *for all* $s = 1, 2, \ldots, S$. *Define* $T' = V_1 + \cdots + V_S$. *Then*

$$Pr(T \geq a | \mathcal{F},\, \mathcal{Z} \cap \mathcal{A}_{\mathbf{1}}) \leq Pr\big(T' \geq a\big) \;\; \textit{for all} \;\; a \geq S\overline{\mathbf{p}}\,.$$

Proof. Observe that $\mathbf{p}$ majorizes $\overline{\mathbf{p}} \cdot \mathbf{1}$ and note that if a function $f(\mathbf{p})$ is Schur-convex in $\mathbf{p}$ then $f(\mathbf{p}) \geq f(\overline{\mathbf{p}}\mathbf{1})$. What remains to be shown is that the distribution function for a Poisson-Binomial is Schur-convex in $\mathbf{p}$. See Gleser (1975) for this approach and Hoeffding (1956) for the original proof. The theorem as stated is an immediate corollary of Theorem 4 in Hoeffding (1956). Gleser (1975) presents a more general version of this result which holds when the success probabilities of T majorize those of T'. □

Remark 1. *Theorem* (2) *is a finite sample result whose proofs we refer to are both rather technical. An analogous asymptotic result follows from much simpler arguments. The variance of a Bernoulli random variable with success* p *can be written as* $f(p) = p(1-p)$. f *is clearly concave and thus by Jensen's Inequality,* $Var(T) \leq Var(T')$. *Since the expectation of* T *and* T' *are equal, using a normal approximation to the exact permutation test will asymptotically yield the same stochastic ordering as in Theorem* (2).

Remark 2. *It is important to note that* $\Gamma'_{truth} \leq \Gamma_{truth}$ *since* $p_s/(1-p_s) \leq \Gamma_{truth}$ *for* $s = 1, \ldots, S$. *Consequently, we have that* $Pr(T' \geq a) \leq Pr(T^+ \geq a)$ *which implies that sensitivity analysis with respect to* Γ', *the average case calibrated sensitivity analysis, is less conservative than the worst case calibrated sensitivity analysis with respect to* Γ.

The implication of this theorem is that it is safe to interpret a sensitivity analysis in terms of Γ', an upper bound on the sample average hidden bias ($\overline{\mathbf{p}}/(1-\overline{\mathbf{p}})$). For example, when using the *most active cellphone day* control window we have $\Gamma_{sens} = 2.4$. Previously, we would say that if no case-crossover pair was subject to hidden bias larger than 2.4, then the data would still provide evidence that talking on a cellphone increases the risk of getting in a car accident. Now, some case-crossover pairs may be subject to hidden bias (much) larger than 2.4, as long as the sample average hidden bias is no larger than 2.4. It is important to note that this interpretation is only valid for binary outcomes. The proof relies on Schur-

convexity of the distribution function of our test statistic with respect to $\mathbf{p}$ which requires that it be symmetric in $\mathbf{p}$. For more general tests, such as the sign-rank test, this is not the case.

Some additional applications of Theorem 2 can be found in the Appendices. Appendix A considers the case when U_{s1} and U_{s1} measure some time-varying propensity of subject s to use his cellphone. Using Theorem 2 we develop a little theory and a numerical example. Appendix B provides details on how Theorem 2 can be applied when U is not restricted to the unit interval.

2.4. The Effect of Cellphone Use on Motor-vehicle Collisions

In this section we return to our motivating example to see how our average case theory can provide interpretive assistance to our standard sensitivity analysis we carried out in Section 2.1.4 and allow us to incorporate additional information to empirically calibrate our average case sensitivity analysis.

2.4.1. Driving intermittency

The study conducted in Tibshirani and Redelmeier (1997) did not have access to direct information on whether an individual was driving during the control window. The authors examine the effect of driving intermittency during the control window on their relative-risk estimate by bootstrapping the estimate using an intermittency rate of $\widehat{\rho} = 0.65$. In other words, they correct for bias due to the possibility that a subject was not driving during the control window. The intermittency rate was estimated using a survey asking 100 people who reported car crashes whether they were driving at the same time the previous day. Alternatively, one may ask a related question in the context of a sensitivity analysis - does the bias due to driving intermittency explain the observed association between cellphone usage and traffic incidents? Given that the study took place in the early 1990s when, for some cellphones and carphones were synonymous, it would not be surprising if many study participants (almost) always used their landlines rather than their cellphones when not

driving, violating the positivity assumption. Therefore, the only plausible Γ_{truth} is infinite (or at least very large) when conditioning on case-crossover pairs where the subject is on her cellphone in only one of the two windows. This renders the worst case sensitivity analysis uninformative. No magnitude of association between cellphone use and car accidents would convince us that the relationship was causal if we stuck to the worst case calibration of the sensitivity analysis. The average case calibration, on the other hand, still has a chance. We can use our estimate $\widehat{\rho}$ to approximate a plausible value of $\overline{\mathbf{p}}$, $\overline{\mathbf{p}} = (1-\widehat{\rho})\cdot 1 + \widehat{\rho}\cdot 0.5 = 0.675$, and a corresponding plausible value of Γ'_{truth}, $\Gamma'_{truth} = \overline{\mathbf{p}}/(1-\overline{\mathbf{p}}) = 2.1$. Theorem (2) circumvents the conceptual hurdle of unbounded Γ_{truth} and allows us to confidently use a sensitivity analysis to quantitatively assess the causal evidence. Moreover, it allows us to incorporate information about ρ into our analysis. If the association between cellphone use and motor vehicle collisions is causal in nature, our empirical calibration suggests that our test for treatment effect should be insensitive to unobserved biases with magnitude $\Gamma' \approx 2.1$.

2.4.2. An alternative approach to handling pairs with unbounded bias

There are other approaches to dealing with the example of infinite bias we just presented. For instance, the investigator may be more confident in specifying an upper bound on the worst case bias to be finite, $\Gamma < \infty$, for a proportion $1-\beta$ of the matched pair sample than he is in working in terms of the average case bias. If he has a good sense of what proportion β of the pairs is exposed to unbounded bias he may drop $\beta \times S$ pairs where the treated unit had positive outcome and perform the standard worst case sensitivity analysis on the remaining $(1-\beta) \times S$ pairs. Rosenbaum (1987) proved that this method yields a valid sensitivity analysis. This strategy would be particularly suited for the example of driver intermittency discussed above. However, this approach assumes this particular pattern of unmeasured confounding is present and driver intermittency is just one of many sources of potential bias. On the other hand, the average case analysis accomodates arbitrary patterns of bias that may lead to large differences in average and worst case biases.

2.4.3. Average case sensitivity analysis for attributable effects

How many of the recorded accidents in our study can be attributed to the driver talking on a cellphone? Recall from Section 2.2.3 that the set indicated by $\boldsymbol{\delta}^*$ includes the a pairs in $D(+,-)$ with the smallest values of p_s. Although we cannot compute $\widetilde{T}_{-\boldsymbol{\delta}^*}$ and thus cannot use it directly to conduct inference, we can compute a lower bound that we will show can be used to perform an average case sensitivity analysis:

$$\begin{aligned}
\widetilde{T}_{-\boldsymbol{\delta}^*} &= \frac{\sum_{s=1}^{S} Z_{s1} r_{Cs1} - \sum_{s=1}^{S}(1-\delta^*_{s1})p_s}{\left\{\sum_{s=1}^{S}(1-\delta^*_{s1})p_s(1-p_s)\right\}^{1/2}} \\
&= \frac{\sum_{s=1}^{S} Z_{s1} R_{s1}(1-\delta^*_{s1}) - \sum_{s=1}^{S}(1-\delta^*_{s1})p_s}{\left\{\sum_{s=1}^{S}(1-\delta^*_{s1})p_s(1-p_s)\right\}^{1/2}} \\
&= \frac{T - a - (S-a)\overline{\mathbf{p}}(a)}{\left\{\sum_{s=1}^{S}(1-\delta^*_{s1})p_s(1-p_s)\right\}^{1/2}} \\
&\geq \frac{T - a - (S-a)\overline{\mathbf{p}}(a)}{\{(S-a)\overline{\mathbf{p}}(a)(1-\overline{\mathbf{p}}(a))\}^{1/2}} = \widetilde{T}(\overline{\mathbf{p}}(a))
\end{aligned} \tag{2.9}$$

where $\overline{\mathbf{p}}(a) = \sum_{s=1}^{S}(1-\delta^*_{s1})p_s/(S-a)$. The last inequality follows from Jensen's inequality applied to the variance term in the denominator. Notice that instead of applying Theorem (2) in order to derive a sensitivity analysis in terms of the average bias we use the simpler argument in Remark (1). Now note that if $p_s \geq p_*$ for all $s = 1, \ldots, S$ then we can relate the trimmed average probability, $\overline{\mathbf{p}}(a)$, to $\overline{\mathbf{p}}$ as follows

$$\overline{\mathbf{p}} \geq \frac{(S-a)\overline{\mathbf{p}}(a) + a \cdot_*}{S} = q(a)\,. \tag{2.10}$$

We can use this relationship to construct a simple procedure – mirroring that of Section 2.2.3 – to perform an average case calibrated sensitivity analysis for one-sided confidence intervals of the form $\{A : A > a\}$ that yields average case calibrated sensitivity intervals. The procedure can be summarized as follows,

1. Choose a desired average calibrated sensitivity parameter Γ'.

2. For $a = 0$ solve $q(a) = \Gamma'/(1 + \Gamma')$ for $\overline{\mathbf{p}}(a)$ and denote the solution $p(a, \gamma')$. Compute $\widetilde{T}(p(a, \gamma'))$.

3. If $\widetilde{T}(p(a, \gamma')) < \Phi^{-1}(1 - \alpha)$ then conclude it is plausible that none of the accidents can be attributed to talking on a cellphone.

4. Else, repeat steps (2) and (3) for $a = 1, \ldots, S$ stopping when $\widetilde{T}(p(a, \gamma')) < \Phi^{-1}(1 - \alpha)$. Let $a^* = a - 1$.

5. Return the $100 \times (1 - \alpha)\%$ sensitivity interval $\{A : A > a^*\}$ and conclude that it is plausible that more than a^* of the accidents are attributable to talking on a cellphone when exposed to an average bias of at most Γ'.

Just as in the simple test for no treatment effect, we see that we have a nearly identical procedure to the worst case sensitivity analysis with an average interpretation of the bias parameter. In fact, the procedure also yields a corresponding worst case calibration for the computed sensitivity interval. Under the worst case calibration, the sensitivity interval from step (5) would correspond to a worst case bias $\Gamma = p(a^*, \gamma')/(1 - p(a^*, \gamma'))$.

How might we apply this procedure to our example? For a given control window we would like to make confidence statements such as, *at the 95% level it is plausible that there are a^* or more accidents attributable to talking on a cellphone.* Recall the empirically calibrated average case bias from Section 2.4.1, $\Gamma' \approx 2.1$. We may also be interested making sensitivity statements such as, *if the average probability of talking on a cellphone during the hazard window is at most 2.1 times that of talking on a cellphone in the control window for drivers in our study, $\Gamma' = 2.1$, it is plausible at the 95% level that there are a^* or more accidents attributable to talking on a cellphone.* Table 4 summarizes the plausible range of attributable accidents for each of the four different control windows. For all four control windows we set $\Gamma' = 2.1$. The first column is the number of discordant pairs in which the driver was on

a cellphone during the control window. The second column reports the lower bound a^* of the one-sided sensitivity intervals for $\alpha = 0.05$ We also report the corresponding worst case calibrated bias in the last column of Table 4. In the cellphone study we have no convincing reason to believe that $p_* > 0$ but in other examples, it may make sense that p_s is bounded from below, which has the effect of making the procedure less conservative.

Control Window	$\lvert D(+,-)\rvert$	a^*	Γ'	Γ
previous weekday/weekend	158	28	2.1	4.04
one week prior	164	31	2.1	4.37
previous driving day	119	18	2.1	3.51
most active cellphone day	134	5	2.1	2.3

Table 4: Sensitivity analysis for 95% one-sided confidence intervals for attributable effects of the form $\{A : A > a^*\}$. Γ' indicates the average calibration bias that we specify for the procedure and Γ is the implied worst case calibration that corresponds to the computed interval. We assume that $p_* = 0$.

We find that even if the average probability of talking on a cellphone during the hazard window was at most 2.1 times that of talking on a cellphone on the same day one week prior, it is plausible that there are 31 or more accidents attributable to talking on a cellphone. The implied worst case bias associated with this statement is $\Gamma = 4.37$. What this means is that we would arrive at the same conclusion about the number of plausible attributable accidents if we put an upper bound on the worst case bias of $\Gamma = 4.37$ and followed the standard confidence interval procedure for attributable effects outlined in Section 2.2.3 and Gastwirth et al. (2000). Unlike the sensitivity analysis for the simple test for no treatment effect, the average case calibrated sensitivity analysis for attributable effects is not guaranteed to be less conservative than the worst case calibration. For a 95% sensitivity interval for attributable effects generated by our procedure where $a^* > 0$, the corresponding upper bound on the average case bias Γ' is less than the corresponding upper bound on the worst case bias Γ. This occurs since we do not know which pairs contain attributable effects nor do we know each pair's particular exposure to hidden bias. Without any further assumptions, the best lower bound for Γ' assumes that all the a pairs with attributable effects have arbitrarily

small probability of being on a cellphone in the hazard window and not the control window. This is expressed mathematically in equation (2.10) by setting $p_* = 0$. If $\Gamma'_{truth} < \Gamma_{truth}$ – which is a reasonable assumption in most circumstances – then using the average case calibration may still result in a less conservative analysis. However, if all case-crossover pairs are exposed to the same magnitude of bias such that $\Gamma'_{truth} = \Gamma_{truth}$ then we are guaranteed to be less conservative by using the worst case calibration. A reasonable solution would be to simply supply both the Γ' and Γ when reporting a sensitivity interval, as we do in Table 4. The investigator may then present an argument based on subject matter expertise as to which calibration is likely to be less conservative.

2.5. Discussion

The theorem presented in 2.3.1 can be thought of as an interpretive aid: For the same standard sensitivity analysis we now have an additional, often more natural, way to interpret the results. This new average case interpretation may also allow researchers to make use of additional information about the problem to empirically calibrate their sensitivity analysis. As we saw in Section 2.4.1, we used the estimate of driver intermittency rate to determine an approximate lower bound on Γ'_{truth}, providing us with some empirical guidance when conducting our sensitivity analysis. In the worst case setting, such an empirical calibration would not be possible. The investigator performs a sensitivity analysis in anticipation of critics who might claim the association is due to some unobserved confounder. The average case analysis makes the protection that the sensitivity analysis provides against such criticism more robust. As the title of the article makes clear, the results we present are for binary data. As we illustrated in Section 2.4.3, the notion of attributable effects allows us to construct interpretable confidence intervals for binary outcomes. We show that our average case calibration can be extended to the sensitivity analysis of such confidence intervals and in most cases will yield a less conservative conclusions. It may then be interesting to apply the results here to the sensitivity analysis of displacement effects, the continuous analog of attributable effects for non-binary outcomes. Rosenbaum (2002a) show that displacement

effects can be analyzed in the attributable effect framework for binary response, providing a potential avenue to extend average case calibrated sensitivity analysis to a study with non-binary outcomes.

2.6. Appendix

2.6.1. Appendix A

Time-varying propensity for cellphone use

After conditioning on the time-invariant driver characteristics $\mathbf{X}$, it may be natural to model U_{si} as a time-varying propensity quantile for using a cellphone in the hazard window ($i = 1$) and in the control window ($i = 2$). U_{si} could summarize an arbitrary number of confounding variables that vary between control and hazard windows for driver s. As a quantile, we can think of U_{si} as coming from a uniform distribution on $[0, 1]$. U_{s1} and U_{s2} can conceivably be considered independent since by design a case-crossover study controls for all individual, time-invariant confounders. Now, suppose that we conduct a sensitivity analysis that returns Γ_{sens}. If we interpret this as an average case hidden bias we may want to ask how large we would expect the corresponding worst case hidden bias to be. If we assume the the U_{si} are propensity quantiles that are iid uniformly distributed we can compute a lower bound for the expected worst case hidden bias corresponding to the average case calibrated Γ_{sens}. Let Σ be the set of all permutations of $\{11, 12, \ldots, S1, S2\}$ and let $\sigma \in \Sigma$ be an element in the set. Now define $\gamma^*(\mathbf{U})$ to be the solution to

$$\Gamma_{sens}/(1 + \Gamma_{sens}) = \sup_{\sigma \in \Sigma} \left\{ \frac{1}{S} \sum_{s=1}^{S} \frac{\exp(\gamma(U_{\sigma(s1)} - U_{\sigma(s2)}))}{1 + \exp(\gamma(U_{\sigma(s1)} - U_{\sigma(s2)}))} \right\}. \tag{2.11}$$

The right hand side of this equation inside the supremum operator is an expression for $\bar{\mathbf{p}}$ under the sensitivity model defined in Section 2.2. The following proposition and corollary show that $\Gamma^* = \exp(\gamma^*(\mathbf{U}))$ is a lower bound for the worst case calibrated hidden bias corresponding to the average case calibrated Γ_{sens} given $\mathbf{U}$.

Proposition 1. *With probability one* $\gamma^*(\mathbf{U})$ *is the unique solution of* (2.11) *and the smallest* γ *that satisfies*

$$\Gamma_{sens}/(1+\Gamma_{sens}) = \frac{1}{S}\sum_{s=1}^{S}\frac{\exp(\gamma(U_{\sigma(s1)} - U_{\sigma(s2)}))}{1+\exp(\gamma(U_{\sigma(s1)} - U_{\sigma(s2)}))}$$

for some $\sigma \in \Sigma$.

Proof. It suffices to show that the right hand side of (2.11) is strictly increasing in γ with probability one. Consider $0 \leq \gamma_1 < \gamma_2$ and let σ_1 be the permutation that maximizes

$$f(\sigma, \gamma_1, \mathbf{U}) = \frac{1}{S}\sum_{s=1}^{S}\frac{\exp(\gamma_1(U_{\sigma(s1)} - U_{\sigma(s2)}))}{1+\exp(\gamma_1(U_{\sigma(s1)} - U_{\sigma(s2)}))}\,.$$

Assuming that the U_{si} are iid uniform, $\mathbf{U}$ is nonconstant with probability one. If $\mathbf{U}$ is nonconstant then $f(\sigma_1, \gamma_1, \mathbf{U}) < f(\sigma_1, \gamma_2, \mathbf{U})$ since $U_{\sigma(s1)} - U_{\sigma(s2)} > 0$ for at least one $s = 1, \ldots, S$. If we let σ_2 be the permutation that maximizes $f(\sigma, \gamma_2, \mathbf{U})$ then we have that $f(\sigma_1, \gamma_1, \mathbf{U}) < f(\sigma_2, \gamma_2, \mathbf{U})$, completing the proof. □

Corollary 3. *Under the sensitivity model defined in Section 2.2,* $\Gamma^* = \exp(\gamma^*(\mathbf{U}))$ *is a lower bound for the worst case calibrated hidden bias corresponding to the average case calibrated* Γ_{sens} *given* $\mathbf{U}$.

Using Corollary 3 we can determine the expected lower bound on the worst case hidden bias corresponding to the average case calibrated Γ_{sens} by computing $\mathbb{E}[\Gamma^*]$ via Monte Carlo estimation. The expectation here is taken over $\mathbf{U} \sim \mathcal{U}[0,1]^{2S}$. Under this propensity quantile model for the unobserved confounders, this procedure can give us a sense of how much less conservative the average case calibration is than the worst case calibration. In the table below we give Monte Carlo estimates and standard errors for $\mathbb{E}[\Gamma^*]$ corresponding to the average case calibrated Γ_{sens} for each of the four control windows.

For each control window we find that the average case interpretation is significantly less conservative than the worst case interpretation. For example, using the worst case calibrated

Control Window	Γ_{sens}	$\mathbb{E}[\Gamma^*]$
previous weekday/weekend	4.92	24.76 (0.08)
one week prior	5.53	31.42 (0.11)
previous driving day	4.15	17.66 (0.06)
most active telephone day	2.40	5.81 (0.01)

Table 5: Sensitivity analysis for (marginal) $\alpha = 0.05$ and the expected lower bound on corresponding worst case calibrated bias $\mathbb{E}[\Gamma^*]$. Standard errors for Monte Carlo estimates are in parentheses.

sensitivity analysis we can conclude that if no case-crossover pair was subject to hidden bias larger than 4.15, there is still significant evidence at level $\alpha = 0.05$ that talking on the phone increases the risk of getting in a car accident. In contrast, using the average case calibration we can say there is significant evidence of a treatment effect even if we *expect* that the worst case bias in any case-crossover pair to be greater than 17.66. Notice that for larger values of Γ_{sens} the benefit from using the average case calibration increases. This exercise is not necessarily meant to be a general purpose procedure but rather a numerical illustration of the gain in power that comes from using the average case calibration even under relatively innocuous assumptions about $\mathbf{U}$.

2.6.2. Appendix B

Simultaneous sensitivity analysis

The one-parameter sensitivity model introduced in Section 2.2 is often referred to as the *primary sensitivity analysis*. In this model, the association between U_{si} and Z_{si} is controlled by Γ and the stochastic ordering in Equation (5) of the main paper were derived by Rosenbaum (1987) assuming that U_{si} and r_{Csi} have a near perfect relationship but this is not always a plausible assumption – for example, if U_{si} is continuous propensity score for treatment and the outcome is binary. A more general two-parameter sensitivity model was first introduced by Gastwirth et al. (1998) where Δ controls the association between U_{si} and r_{Csi} and Λ controls the association between U_{si} and Z_{si}. For example, if $U_{s1} = 1$ and

$U_{s2} = 0$ then the first unit of the pair is Λ times more likely to receive treatment and Δ times more likely to have the positive outcome. This model is known as the *simultaneous sensitivity analysis* and is particularly useful when it is not plausible that U_{si} and r_{Csi} are perfectly correlated. When the outcome is binary – and more generally if y_s, the difference in outcomes in pair s, come from a distribution belonging to Wolfe's semiparametric family (see Wolfe (1974)) – Rosenbaum and Silber (2009) show that Γ can be *amplified* to the two-parameter model (Λ, Δ) by the identity $\Gamma = (\Lambda\Delta + 1)/(\Lambda + \Delta)$ which we refer to as the *amplification curve*. The simultaneous sensitivity model acts as an interpretive aid to the standard one-parameter procedure; We may consider how u_{si} affects the odds of treatment and the odds of positive outcome separately and then use the amplification curve to determine the corresponding Γ with which we can perform a standard one-parameter sensitivity analysis. Like the one-parameter sensitivity model, the simultaneous sensitivity model does not require that the investigator specifies the distribution of U_{si}, only that $U_{si} \in [0, 1]$ for $s = 1, 2, \ldots, S$ and $i = 1, 2$.

When U is not bounded

Theorem 1 in the main paper is free from any modeling decision of the underlying causal mechanism. In that sense, it is very general and allows us to relax the restriction that U lie in the unit interval. This provides flexibility in modeling the unobserved confounders but for Γ, or Δ and Λ in the amplified setting, to retain meaning we will have to standardize the distribution of U in some fashion. For example, we may scale U such that the post matching variance of $U_{s1} - U_{s2}$ is equal to 1. Now suppose that $U_{s1} - U_{s2} = \pm 1$. Then the odds that the treated unit has positive outcome in pair s is

$$\Gamma = \frac{1 + \Lambda\Delta}{(1 + \Lambda)(1 + \Delta)}, \tag{2.12}$$

which is equivalent to the worst case bias when U was taken to lie on the unit interval. We will refer to this as the one standard deviation (1SD) worst case bias. Related work

by Wang and Krieger (2006) considers arbitrary distributions of $W_s = U_{s1} - U_{s2}$ with mean 0 and variance 1 after matching. They show that for any such distribution of $\mathbf{W} = (W_1, W_2, \ldots, W_S)^T$, the population mean of p_s, $\mathbb{E}[p_s]$, is maximized when W_s takes values ± 1 with equal probability. The implication of this result is that when U is scaled appropriately the 1SD worst case bias is asymptotically more conservative than Γ' even when U is not restricted to the unit interval.

CHAPTER 3

Extended Sensitivity Analysis for Heterogeneous Unmeasured Confounding with an Application to Sibling Studies of Returns to Education

Abstract

The conventional model for assessing insensitivity to hidden bias in paired observational studies constructs a worst-case distribution for treatment assignments subject to bounds on the maximal bias to which any given pair is subjected. In studies where rare cases of extreme hidden bias are suspected, the maximal bias may be substantially larger than the typical bias across pairs, such that a correctly specified bound on the maximal bias would yield an unduly pessimistic perception of the study's robustness to hidden bias. We present an extended sensitivity analysis which allows researchers to simultaneously bound the maximal and typical bias perturbing the pairs under investigation while maintaining the desired Type I error rate. We motivate and illustrate our method with two sibling studies on the impact of schooling on earnings, one containing information of cognitive ability of siblings and the other not. Cognitive ability, clearly influential of both earnings and degree of schooling, is likely similar between members of most sibling pairs yet could, conceivably, vary drastically for some siblings. The method is straightforward to implement, simply requiring the solution to a quadratic program.

3.1. Introduction

3.1.1. A motivating example: Returns to schooling

Is educational attainment a determining factor for success in the labor market? Initial interest among economists in addressing this question is attributed to the observation in the late 1950s that increases in education levels could account for much of the productivity growth in post-war US (Becker, 2009; Griliches, 1970; Card, 1999). With strong evidence of a positive association between education and earnings in a variety of political and geographic environments but little to no experimental data, a recurring theme in the

subsequent pursuit of a causal relationship between education and income is that of the presence of "ability bias" (Card, 1999). After controlling for family background, or considering within-family estimates of the causal effect using sibling or twin studies, can latent differences in ability influence both differences in schooling choice and earnings? A notable twin study by Ashenfelter and Rouse (1998), which we re-examine in this paper, argued cogently, albeit with limited statistical evidence, that identical twins can be regarded as truly identical in all dimensions relevant to schooling choices and future income, including latent ability. In a survey of contemporary economic investigations of returns to education, Card (1999, p.1852) addresses this hypothesis:

> Despite this evidence, and the strong intuitive appeal of the "equal abilities" assumption for identical twins, however, I suspect that observers with a strong a priori belief in the importance of ability bias will remain unconvinced.

Perhaps latent ability is truly identical for many twin pairs but markedly different in a few pairs; what would happen then? That exogeneity is not testable leaves even the most compelling observational evidence susceptible to the warranted, though often non-specific, criticism, "what if bias remains?" Should the totality of evidence assume the absence of hidden bias, the critic need merely suggest the existence of bias to cast doubt upon the posited causal mechanism. It is thus incumbent upon researchers not only to anticipate such criticism, but also to arm themselves with a suitable rejoinder. Rather than arguing for or against the presence of ability bias or any other unobserved confounding factor, in this paper we assess the sensitivity of causal conclusions to departures from truly randomized assignment while allowing for patterns of ability bias that may be highly heterogeneous across sibling pairs.

3.1.2. Assessing returns to schooling with sibling comparison designs

Sibling comparison studies are a special case of stratified designs where natural blocks are formed by family membership. These studies automatically control for genetic, socioeco-

nomic, cultural, and child-rearing characteristics to the extent that they are shared between siblings; however, instability of familial characteristics over time for sibling pairs of different ages and non-shared genetic makeup are among threats to this premise (Donovan and Susser, 2011). Due to their natural and automatic control of stable familial factors, both observed and unobserved, sibling comparison designs have long been a popular tool for studying causal effects in both epidemiological and economic settings; see Griliches (1979) and Donovan and Susser (2011) for surveys of past and current sibling comparison studies in economics and epidemiology, respectively.

Sibling comparison designs have been particularly fruitful in the study of returns to schooling, where genetic and family background are deemed essential to both schooling choices and future income; see for example Hauser et al. (1999), Stanek et al. (2011), and Ashenfelter and Rouse (1998). Hauser et al. (1999) study sibling pairs from the Wisconsin Longitudinal Study (WLS), a random sample ($n = 10,317$) of men and women born between 1938 and 1940 who graduated from Wisconsin high schools in 1957. The size of the sample was set to be approximately a third of all Wisconsin high school graduates in 1957. Random siblings of those in the study ($n = 7,928$), born between 1930 and 1948, were also selected and interviewed. The WLS contains a rich set of baseline covariates and endpoints, including physical, cognitive, social, and occupational outcomes collected over nearly 60 years following graduation. Uniquely, the WLS dataset contains intelligence quotient (IQ) scores recorded while a given individual was in high school – a covariate rarely measured in longitudinal cohort studies.

In other sibling studies of the returns to schooling, such as that of Ashenfelter and Rouse (1998), baseline intelligence measures such as IQ are not available, making it plausible that the siblings being compared differ in cognitive ability in unobserved ways. Furthermore, the IQ data from the WLS study suggests that, when considering same-sex sibling pairs where one sibling attended college and the other did not ($n = 171$), intellectual ability is not balanced sufficiently by shared genetics alone. The boxplots of differences in IQ between

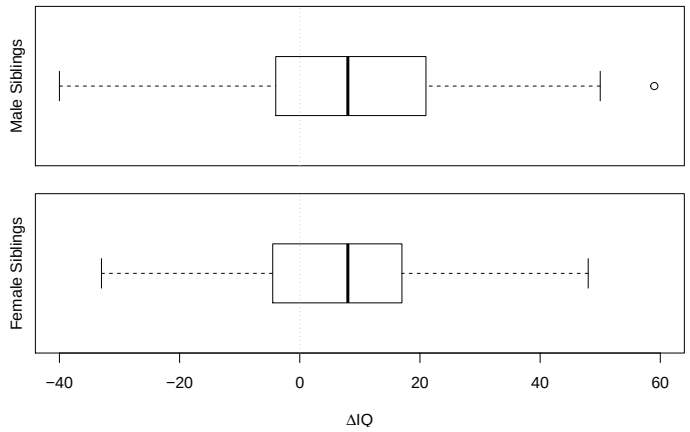


Figure 1: Boxplots of differences in IQ scores between same-sex siblings where one attended college and the other did not. *(top panel):* Male same-sex sibling pairs ($n = 128$). *(bottom panel):* Female same-sex sibling pairs ($n = 43$).

the college-attending siblings and their counterparts in Figure 1 exhibit a prominent shift in the IQ distribution between the two groups for both male and female same-sex sibling pairs. The mean (sd) is 107.1 (14.7) in the college-attending group and 97.4 (14.4) in the high school-only group for male same-sex sibling pairs. In female same-sex sibling pairs, these values are 108.1 (14.0) and 101.4 (14.2) for the college-attending and high school-only attending groups respectively. Details on the construction of the 171 same-sex sibling pairs can be found in Appendix 3.8.2. An important inclusion criterion was that both siblings were employed when income data was collected.

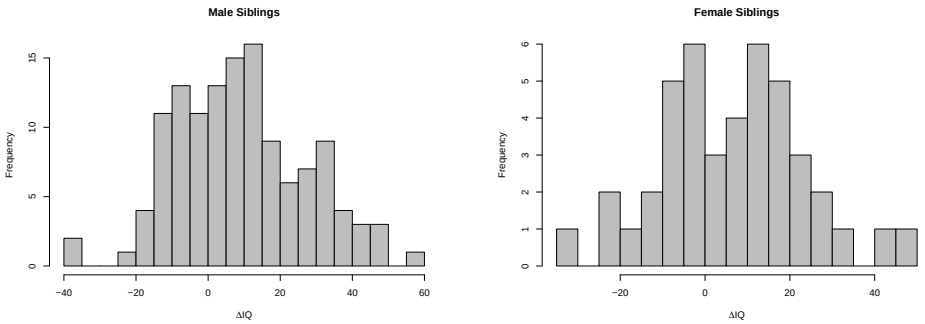
3.1.3. Potential for rare but extreme unmeasured biases

Despite their analytical strengths and convenient, automatic stratification, sibling comparison designs for estimating causal effects are subject to biases arising from differences in subject-level confounders. For example, latent ability, as measured by IQ, may differ substantially within twin pairs in Ashenfelter and Rouse's twin study. This concern is magnified in sibling studies where discordant within-pair treatment assignment may actually exacerbate differences in covariates that are related to both the intervention and outcome of interest (Frisell et al., 2012). When pairs do not arise naturally, as in paired sibling studies, matching algorithms designed to minimize disparities in observed covariates may

be used to construct pairs of "comparable" subjects; see, for example, Hansen and Klopfer (2006) and Stuart (2010) for discussion on various approaches to matching. Matched pairs constructed in this fashion may be comparable along observed covariates, but they are still vulnerable to unmeasured bias arising from differences in covariates not available to the matching algorithm.

While agnostic covariate adjustment within sibling sets as suggested in Rosenbaum (2002b) can help mitigate the impact of discrepancies in observed individual-specific covariates, bias arising from differences in unobserved confounders may remain and imperil the conclusions of the study. An additional inferential step known as a *sensitivity analysis* assesses the robustness of the conclusions of a study to these unmeasured biases. Sensitivity analysis was first introduced by Cornfield et al. (1959) and refined to accommodate continuous outcomes in Rosenbaum (1987). The resulting sensitivity analysis for paired studies considers the worst-case bias to which any pair may be subject and asks whether the study conclusions might change if we assumed that *all* pairs were exposed to the maximal bias in a manner adverse to the desired inference. We refer to this as the *conventional* sensitivity analysis. See Cornfield et al. (1959), Marcus (1997), Imbens (2003), Yu and Gastwirth (2005), Wang and Krieger (2006), Egleston et al. (2009), Hosman et al. (2010), Zubizarreta et al. (2013), Liu et al. (2013), and VanderWeele and Ding (2017) for additional perspectives on and worked examples of sensitivity analysis.

In many paired studies, sibling or otherwise, hidden biases may strongly influence the results observed for some pairs and more modestly affect others. If the impact of unmeasured confounding were truly heterogeneous in this manner, the conventional sensitivity analysis would be conspicuously conservative. Consider, for example, discrepancies in IQ scores within sibling pairs measured in the WLS where one sibling attended college for at least two years and the other received at most a high school diploma. While existing longitudinal cohort studies rarely contain measures of intelligence (Herd et al., 2014), existing evidence suggests that discrepancies in IQ between sibling pairs are strongly predictive of both dif-



Odds Ratio	$[1,2)$	$[2,3)$	$[3,6)$	$[6,7)$	$[7,9)$	$[9,10)$
Count	165	4	0	1	0	1

Figure 2: *(left panel)*: Histogram of between-sibling IQ disparities of same-sex male sibling pairs in the WLS study where one sibling attended college and the other did not ($n = 128$). *(right panel):* Histogram of between-sibling IQ disparities of same-sex female sibling pairs in the WLS study where one sibling attended college and the other did not ($n = 43$). *(bottom panel)*: Table of the estimated increase in pairwise bias due to IQ disparities between siblings measured as an odds ratio.

ferences in educational attainment and differences in future income (Stanek et al., 2011). In the WLS data, the between-sibling disparity in IQ scores is quite variable across sibling pairs where one sibling attended college and the other did not. The histogram of these college-minus-high school differences is shown in the left panel of Figure 2 for male sibling pairs and the right panel for female sibling pairs. Most IQ differences are modest, but a few sibling pairs have large imbalances (e.g. > 40).

In a sibling study on the returns of schooling where IQ was not recorded, such as Ashenfelter and Rouse's twin study, the maximal bias to which any pair is subject could be materially larger than the typical bias for any sibling pair. Evidence of this pattern's plausibility can be seen in the bottom table of Figure 2. The table shows the distribution of the estimated increase in pairwise bias due to IQ disparities between siblings measured as an odds ratio. The numerator of the odds ratio is the predicted maximum odds that the sibling who reported higher income attended college given the reported disparities in IQ while the denominator corresponds to the maximum odds had both siblings had the same IQ. (the method for estimating these odds ratios is described in Appendices 3.8.3-3.8.4). While the

odds ratio in most pairs is close to one, there are a handful of pairs with odds ratios near 2 and two rare cases of odds ratios greater than 6. As far as the 'typical' or 'expected' pairwise bias is as interpretable a quantity as the worst-case pairwise bias, an *extended* sensitivity analysis of both maximal and expected bias may alleviate concerns that the conventional approach is overly pessimistic while providing a more flexible handling of unobserved bias.

3.1.4. Accommodating varying degrees of unmeasured confounding

We present an extended sensitivity analysis bounding both the maximal and expected bias for paired studies. The concept of expected bias is made precise in §3.3.1. The theoretical foundations and implementation of the extended sensitivity analysis are developed in §§3.2-3.4, while supporting Type I error control and power simulations are presented in §3.5. The procedure involves two interpretable parameters, Γ and $\bar{\Gamma} \leq \Gamma$, bounding the maximal and expected bias, respectively. At one extreme, setting $\bar{\Gamma} = \Gamma$ recovers the conventional sensitivity analysis for paired studies proposed in Rosenbaum (1987, §2). At the other, setting $\Gamma = \infty$ for a fixed value of $\bar{\Gamma}$ allows one to bound the average bias while leaving the maximal bias in any given pair unbounded, subsuming the extension presented in Rosenbaum (1987, §4) where the investigator specifies a fraction β of the pairs that satisfy a constraint on the maximal bias and allows the remaining pairs to be exposed to potentially unbounded bias.

The procedure builds on recent work by Hasegawa and Small (2017) that established an exact sensitivity analysis for the sample average bias for paired studies with binary outcomes in two important ways. First, our procedure accommodates continuous outcomes while providing an asymptotically valid testing procedure for sharp null hypotheses for a large class of test statistics. While generalizing to continuous outcomes corrupts properties unique to McNemar's test statistic utilized in Hasegawa and Small (2017), these difficulties are overcome through a new formulation of the optimization problem necessitated by the sensitivity analysis as a quadratic program. Second, our procedure allows the researcher to bound the expected bias at the level of a superpopulation, rather than the average of the bias at the level of the observed study population, if a superpopulation model is deemed ap-

propriate. This facilitates consonance between superpopulation and finite-sample modes of inference to which the researcher is automatically entitled when only bounding the maximal bias. Actualizing this harmony requires the combination of concentration inequalities with the technique presented in Berger and Boos (1994) for yielding valid p-values by maximizing over a confidence set for nuisance parameters.

To demonstrate the practical consequences of our procedure we return in §3.6 to the motivating example of returns to schooling. Using the availability of IQ measures in the WLS sibling data, we follow Hsu and Small (2013) to estimate the maximal and expected bias under the assumption that inherent cognitive ability is the overwhelming unobserved confounding factor in sibling studies of returns to schooling when IQ measures are not available. We compare standard and extended sensitivity analyses calibrated to these estimates of the sensitivity parameters for Ashenfelter and Rouse's twin study where IQ was not observed.

3.2. Sensitivity analysis for paired studies

3.2.1. An idealized construction of a paired observational study

There are I pairs of individuals. In the i^{th} matched pair one individual receives the treatment, $Z_{ij} = 1$, and the other receives the control, $Z_{ij'} = 0$, such that $Z_{i1} + Z_{i2} = 1$ for each i. In practice, the I pairs come into being by minimizing a metric reflective of the within-pair discrepancies between the observed covariates $\mathbf{x}_{ij}$ for the treated and control individuals in a candidate pairing, such that $\mathbf{x}_{i1} \approx \mathbf{x}_{i2}$ in the resulting pairs. As an idealization of this practice, we follow Rosenbaum (1987) and imagine a generative model where the pairs are constructed, for $i = 1, ..., I$, by initially drawing, without replacement from an infinite population of treated individuals (that is, conditional upon $Z = 1$), an individual who has an observed covariate $X_i = x_i$. For each i, we then sample a control individual from the population of controls with the same value for the observed covariate, i.e. given $Z = 0, X = x_i$. Finally, randomly assign indices $(i, 1)$ and $(i, 2)$ to the two individuals in pair i, and let X_i be a random variable denoting the shared value $X_{i1} = X_{i2}$. Despite having

a shared value X_i, it may be the case that $U_{i1} \neq U_{i2}$ in any pair i for some unobserved covariate U. In §3.3.3, we describe the extent to which the following methodology applies to finite-sample inference in the absence of a superpopulation.

Under the stable unit-treatment value assumption (Rubin, 1980), individual j in matched set i has a potential outcome under treatment, R_{Tij}, and under control, R_{Cij} which does not depend on the treatment received by other individuals in the population. The fundamental problem of causal inference is that vector (R_{Tij}, R_{Cij}) is not jointly observable. Instead, we observe the response $R_{ij} = R_{Tij}Z_{ij} + R_{Cij}(1 - Z_{ij})$, and the observed treated-minus-control paired differences $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$. Lowercase letters denote realizations of random variables. Let $\mathcal{F}_I = \{(x_{ij}, u_{ij}, r_{Tij}, r_{Cij}),\ \ 1 \leq i \leq I,\ \ j = 1, 2\}$ be the values of the potential outcomes, measured covariates, and unmeasured covariates for the $2I$ individuals in the observational study at hand. At times it will be convenient to use boldface for vector-valued constants and random variables after the assignment of indices. For example, $\mathbf{Z}$ represents a vector of length $2I$ with elements $\mathbf{Z} = (Z_{11}, Z_{12}, ..., Z_{I2})$, while $\mathbf{R}_i$ is a vector of length two with elements $\mathbf{R}_i = (R_{i1}, R_{i2})$.

3.2.2. Randomization inference under strong ignorability

The expectation of each paired difference Y_i in the infinite population model of the preceding section is $\mathbb{E}(Y_i \mid X_{ij} = x) = \mathbb{E}(R_{Tij} \mid Z_{ij} = 1, X_{ij} = x) - \mathbb{E}(R_{Cij} \mid Z_{ij} = 0, X_{ij} = x)$ which need not equal $\tau(x) := \mathbb{E}(R_{Tij} - R_{Cij} \mid X_{ij} = x)$ without further assumptions on the relationship between the potential outcomes, the observed covariates, and the treatment indicators. A sufficient condition for equality of these expectations, strong ignorability, entails that for any point x,

$$(R_T, R_C) \perp\!\!\!\perp Z \mid X,\ \ 0 < \mathbb{P}(Z = 1 \mid X = x) < 1. \tag{3.1}$$

Strong ignorability facilitates far more than equality between $\mathbb{E}(Y_i \mid X_{ij} = x)$ and $\tau(x)$; indeed, it entitles the researcher to use randomization tests akin to those justified in ran-

domized experiments. We consider general hypotheses of the form

$$H_0: \quad F_T(R_{Tij}) = F_C(R_{Cij}) \quad \forall i, j$$

for pre-specified functions $F_T(\cdot)$ and $F_C(\cdot)$. While this form accommodates flexible models for treatment effects, perhaps the most classical specification is the additive treatment effect model where the treatment effect is constant at τ for all individuals. Under this model $R_{Tij} = R_{Cij} + \tau$, which can be expressed by setting $F_T(R_{Tij}) = R_{Tij} - \tau$ and $F_C(R_{Cij}) = R_{Cij}$. From our data alone we observe $F_{ij} = F_T(R_{Tij})Z_{ij} + F_C(R_{Cij})(1 - Z_{ij})$; let $\mathbf{F} = [F_{11}, ..., F_{I2}]$. Under H_0, the vectors $\mathbf{F}_C = [F_C(R_{C11}), ..., F_C(R_{CI2})]$ and $\mathbf{F}_T = [F_T(R_{T11}), ..., F_T(R_{TI2})]$ are known to be equal, and hence are entirely specified by the vector of observed responses $\mathbf{R}$.

Let $t(\mathbf{Z}, \mathbf{F})$ be an arbitrary test statistic that is a function of the treatment indicators Z_{ij} and the observed values F_{ij}, and let $\Omega_I = \{\mathbf{z} : z_{i1} + z_{i2} = 1, \quad 1 \leq i \leq I\}$ be the set of 2^I possible assignments of individuals to treatment and control in a paired design. Further let $\mathbf{f}_C$ be the realized value of the random variable $\mathbf{F}_C$. When H_0 holds, $\mathbf{f}_C$ is fully observed. Under the idealized model in §3.2.1 and under (3.1), Theorem 1 of Rosenbaum (1984) demonstrates that under the null hypothesis H_0,

$$\mathbb{P}\{t(\mathbf{Z}, \mathbf{F}) \geq a \mid \mathcal{F}_I, H_0\} = \frac{1}{2^I} \sum_{\mathbf{z} \in \Omega_I} \chi\{t(\mathbf{z}, \mathbf{f}_C) \geq a\}, \tag{3.2}$$

where $\chi\{A\}$ is an indicator that the event A occurred. Importantly, under H_0, the randomization distribution (3.2) is free of unknown parameters through conditioning on $\mathcal{F}_I$, and hence can be used directly to facilitate inference on H_0.

3.2.3. Sensitivity analysis bounding the supremum

In paired randomized experiments, the physical act of randomization breaks the association between potential outcomes and the intervention and thus justifies both the assumption

of strong ignorability and randomization inference through the conditional distribution in (3.2). Paired observational studies aim to mimic an idealized randomized experiment by creating pairs where individuals are similar on the basis of their observed covariates, X, which would similarly facilitate randomization inference through (3.2) if strong ignorability held. In observational studies, strong ignorability, and in turn belief in (3.2), turns a statement of fact into a leap of faith due to the potential presence of unobserved factor U. That treatment assignment is rarely known to be strongly ignorable given observed covariates X alone necessitates a sensitvity analysis which assesses the robustness of a study's conclusions to factors not included in X. A sensitivity analysis operates under the premise that strong ignorability would have been satisfied if an additional pretreatment covariate U had been used in constructing the pairs, that is if for any x and u

$$(R_T, R_C) \perp\!\!\!\perp Z \mid (X, U), \;\; 0 < \mathbb{P}(Z = 1 \mid X = x, U = u) < 1. \tag{3.3}$$

A simple model parameterizing the impact of hidden bias presented in Rosenbaum (1987, §2) relates U to the assignment mechanism through a parameter $\Gamma = \exp(\gamma) \geq 1$, which constrains the degree to which U can affect the odds of receiving the intervention through a logit model,

$$\text{logit}(\mathbb{P}(Z = 1 \mid X = x, U = u)) = \kappa(x) + \gamma u, \;\; 0 \leq u \leq 1. \tag{3.4}$$

The bounds on u in (3.4) may be viewed as a restriction on the scale of the unobserved covariate that is required for the numerical value of γ to have meaning (Rosenbaum, 2002c, Chapter 4). Letting $\pi_i = \mathbb{P}(Z_{i1} = 1 \mid \mathcal{F}_I)$, (3.3) and (3.4) then imply $\pi_i = \text{expit}(\gamma(u_{i1} - u_{i2}))$ and $1 - \pi_i = \text{expit}(\gamma(u_{i2} - u_{i1}))$. As a result, the model requires that the bound $\pi_i^* = \max\{\pi_i, 1 - \pi_i\} = \text{expit}(\gamma|u_{i1} - u_{i2}|) \leq \Gamma/(1 + \Gamma)$ holds uniformly for all i, but imposes no additional constraints on $\boldsymbol{\pi}$, and imposes no constraint on the relationship between the unobserved covariate and the potential outcomes. Theorem 1 of Rosenbaum (1987) illustrates that (3.3), (3.4) and the generative model described in §3.2.1 imply that under

a sharp null H_0, the distribution $t(\mathbf{Z}, \mathbf{F})$ given $\mathcal{F}_I$ takes on the modified form

$$\mathbb{P}\{t(\mathbf{Z}, \mathbf{F}) \geq a \mid \mathcal{F}_I, H_0\} = \sum_{\mathbf{z} \in \Omega_I} \Bigg[\chi\{t(\mathbf{z}, \mathbf{f}_C) \geq a\} \times \prod_{i=1}^{I} \text{expit}(\gamma(u_{i1} - u_{i2}))^{z_{i1}} \text{expit}(\gamma(u_{i2} - u_{i1}))^{z_{i2}} \Bigg]. \quad (3.5)$$

At $\Gamma = 1 \Leftrightarrow \gamma = 0$, (3.5) recovers (3.2), hence representing strong ignorability on the basis of X alone. For $\Gamma > 1$, (3.5) depends on the unknown values of $\mathbf{u}$. A sensitivity analysis proceeds by, for a given value of Γ, finding bounds on (3.5) by optimizing over the nuisance parameters $\mathbf{u} \in [0, 1]^{2I}$ (or equivalently, optimizing over π_i subject to $\pi_i^* \leq \Gamma/(1 + \Gamma)$).

We consider test statistics of the form $t(\mathbf{Z}, \mathbf{F}) = \mathbf{Z}^T \mathbf{q}$ for some function $\mathbf{q} = \mathbf{q}(\mathbf{F})$, commonly referred to as sum statistics. Examples of sum statistics in paired observational studies include Wilcoxon's signed rank test and McNemar's test among many others; see Rosenbaum (2002c, Chapter 2) for more on sum statistics. For example, were we to test the null that the treatment effect was constant at zero for all individuals (commonly referred to as Fisher's sharp null hypothesis), then a choice of $q_{ij} = (R_{ij} - R_{ij'})/I = (r_{Cij} - r_{Cij'})/I$ would amount to a choice of the average of the treated-minus-control paired differences in outcomes as the test statistic. In paired studies, arguments parallel to those in Rosenbaum (2002c, Chapter 4) yield that a tight lower bound on (3.5) is found by setting $u_{i1} - u_{i2} = -\text{sign}(q_{i1} - q_{i2})$ for each pair i, where $\text{sign}(a)$ is the sign of the scalar a. Similarly, a tight upper bound on (3.5) is found by setting $u_{i1} - u_{i2} = \text{sign}(q_{i1} - q_{i2})$ for each i. As a further illustration, if one uses the difference in means as the test statistic, the lower (upper) bound is attained through a perfect negative (positive) correlation between the differences in unmeasured covariates and the signs of the treated-minus-control paired differences.

3.3. An extended sensitivity analysis

3.3.1. Average-case unmeasured confounding in paired studies

In §§1.1-1.2, it was argued that large discrepancies in IQ within pairs of siblings, while likely uncommon, would have a large impact on both likelihood of attaining more than a high school degree and on an individual's expected earnings. Were this the only unmeasured confounder, we would then expect most of the values for $\boldsymbol{\pi}^*$, the maximal probabilities of assignment to treatment within a pair, to not deviate substantially from 0.5, while a few pairs would likely have values for π_i^* substantially larger than 0.5. The conventional model for a sensitivity analysis presented in §3.2.3 bounds π_i^* by $\Gamma/(1+\Gamma)$ for all pairs. Despite typical discrepancies in IQ likely being small, the smallest value of Γ for which (3.4) and (3.5) hold would be large due to the small number of extremely biased pairs. When utilized in its original form, the sensitivity analysis in §2.3 may then paint an overly pessimistic picture of the robustness of the study's findings to unmeasured confounding under this belief, as it cannot account for the 'typical' level of unmeasured confounding being different from the worst-case level.

We consider an extension of the conventional sensitivity analysis summarized in §2.3 involving two sensitivity parameters, Γ and $\bar{\Gamma}$. The first, Γ, plays a role identical to that of Γ in the conventional sensitivity analysis by bounding the supremum of the biased assignment probabilities within a pair. Explicitly, we bound the probabilities of receiving the intervention through a logit form,

$$\text{logit}(\mathbb{P}(Z = 1 \mid X, U)) = \kappa(X) + \gamma U, \;\; 0 \leq U \leq 1. \tag{3.6}$$

That $0 \leq U \leq 1$ trivially implies that for any pair i

$$1/2 \leq \text{expit}(\gamma |U_{i1} - U_{i2}|) \leq \frac{\Gamma}{1+\Gamma}. \tag{3.7}$$

Under (3.3) and the setup of §3.2.1, (3.6) yields that $\Pi_i^* = \max\{\Pi_i, 1-\Pi_i\} = \text{expit}(\gamma|U_{i1} - U_{i2}|) \leq \Gamma/(1+\Gamma)$, where $\Pi_i = \mathbb{P}(Z_{i1} = 1 \mid X_i, \mathbf{U}_i, \mathbf{R}_{Ti}, \mathbf{R}_{Ci}) = \mathbb{P}(Z_{i1} = 1 \mid X_i, \mathbf{U}_i)$. We capitalize U_{ij} and Π_i^* to emphasize that they themselves are random variables with respect to the superpopulation model in §2.1, which would become deterministic by conditioning in $\mathcal{F}_I$.

The second sensitivity parameter, $\bar{\Gamma}$, serves to bound the *expectation* of the biased probabilities. We define $\mu_{\pi^*} = \mathbb{E}[\Pi_i^*] = \mathbb{E}[\text{expit}(\gamma|U_{i1} - U_{i2}|)]$, and impose that for some value $\bar{\Gamma}$ such that $1 \leq \bar{\Gamma} \leq \Gamma$,

$$1/2 \leq \mu_{\pi^*} \leq \frac{\bar{\Gamma}}{1+\bar{\Gamma}}. \tag{3.8}$$

Again, this expectation is taken over repeated samples in the idealized setting in §3.2.1, within which the fixed but unknown values π_i^* in our observational study can be modeled as *iid* realizations of the random variables Π_i^*. As with the conventional sensitivity analysis, our model makes no assumption about the relationship between the unobserved covariates and the potential outcomes.

Like the conventional sensitivity analysis, our extended procedure solves an optimization problem over a set of nuisance parameters $\boldsymbol{\pi}$ that satisfy the typical and maximal bias bounds specified in (3.7) and (3.8). Although the population-level bound $\Pi_i^* \leq \Gamma/(1+\Gamma)$ implies the corresponding sample level bound $\pi_i^* \leq \Gamma/(1+\Gamma)$, the same cannot be said about the corresponding bound on μ_π^*. If $\mu_\pi^* \leq \bar{\Gamma}/(1+\bar{\Gamma})$, a sample realization $\bar{\pi}^*$ arbitrarily close to $\Gamma/(1+\Gamma)$ is still possible, however unlikely. To address this, we translate the bound on μ_π^* to a stochastic bound on $\bar{\Pi}^*$.

In order to construct this stochastic bound, we consider properties of the random variable Π_i^* across draws from the idealized setting in §3.2.1. From (3.7) and (3.8), we have that for all i Π_i^* is bounded above by $\Gamma/(1+\Gamma)$, bounded below by $1/2$, and has expectation μ_{π^*} which is itself bounded above by $\bar{\Gamma}/(1+\bar{\Gamma})$. The Bhatia-Davis inequality (Bhatia and

Davis, 2000) provides the variance upper bound

$$\text{var}(\Pi_i^*) \leq (\Gamma/(1+\Gamma) - \mu_{\pi^*})\,(\mu_{\pi^*} - 1/2) = \nu^2(\Gamma, \mu_{\pi^*}).$$

As the Π_i^* can further be modeled as *iid* random variables under the setting being considered, defining $\bar{\Pi}^* = I^{-1}\sum_{i=1}^{I}\Pi_i^*$, it follows that

$$\mathbb{E}[\bar{\Pi}^*] = \mu_{\pi^*}, \;\; \text{var}(\bar{\Pi}^*) \leq \nu^2(\Gamma, \mu_{\pi^*})/I.$$

If $\text{var}(\Pi_i^*) > 0$ the Central Limit Theorem applies to $\bar{\Pi}^*$, indicating that for any $0 < \beta \leq 0.5$

$$\lim_{I\to\infty} \mathbb{P}(\bar{\Pi}^* \in \mathcal{C}_\beta(\Gamma, \mu_{\pi^*})) \geq 1 - \beta, \tag{3.9}$$

where, because $\bar{\Pi}^* \geq 1/2$ by definition of Π_i^*

$$\mathcal{C}_\beta(\Gamma, \mu_{\pi^*}) = \left[1/2, \mu_{\pi^*} + I^{-1/2}\Phi^{-1}(1-\beta)\nu(\Gamma, \mu_{\pi^*})\right], \tag{3.10}$$

and $\Phi^{-1}(p)$ is the p-quantile of the standard normal distribution. Further, (3.9) is trivially true if $\text{var}(\Pi_i^*) = 0$, as the upper bound of $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$ is no smaller than μ_{π^*} when $\beta \leq 0.5$. That is, knowledge of μ_{π^*} alone enables the construction of asymptotically valid uncertainty sets for $\bar{\Pi}^*$.

3.3.2. Sensitivity analysis bounding the supremum and expectation

Conditional upon $\mathcal{F}_I$, attention returns to the unmeasured confounders for the individuals in our study population, $\mathbf{u}$, and the corresponding assignment probabilities $\boldsymbol{\pi}$. For any value of $\mathbf{u}$ and value for Γ, we have that

$$\mathbb{P}\{t(\mathbf{Z}, \mathbf{F}) \geq a \mid \mathcal{F}_I, H_0\} = \sum_{\mathbf{z}\in\Omega_I} \chi\{t(\mathbf{z}, \mathbf{f}_C) \geq a\} \prod_{i=1}^{I} \pi_i^{z_{i1}}(1-\pi_i)^{z_{i2}}, \tag{3.11}$$

where $\pi_i = \text{expit}(\gamma(u_{i1} - u_{i2}))$. As the shared notation seeks to emphasize, (3.11) is precisely the null distribution utilized in (3.5). Here as well as in (3.5), the unmeasured confounders $\mathbf{u}$, and hence the conditional assignment probabilities $\boldsymbol{\pi}$, are unknown constants, hindering the desired inference through their presence as nuisance parameters. The approach taken in §3.2.3 was to maximize or minimize (3.11) over $\mathbf{u} \in [0,1]^{2I}$ for a given value Γ, or equivalently over $\pi_i^* \leq \Gamma/(1+\Gamma)$. In what follows, we replace this optimization with one over a subset informed by both Γ and $\bar{\Gamma}$ while providing an asymptotically valid level-α test.

Suppose without loss of generality that we are considering a one-sided, greater than alternative. Let $\mathcal{P}_\beta(\Gamma, \mu_{\pi^*}) = \{\boldsymbol{\pi} : \bar{\pi}^* \in \mathcal{C}_\beta(\Gamma, \mu_{\pi^*}),\ \ \pi_i^* \leq \Gamma/(1+\Gamma),\ \ 1 \leq i \leq I\}$, and consider the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{\pi}, \mu_{\pi^*}}{\text{maximize}} \quad & p(\boldsymbol{\pi}, \mu_{\pi^*}) = \sum_{\mathbf{z} \in \Omega_I} \chi\{t(\mathbf{z}, \mathbf{f}_C) \geq t(\mathbf{Z}, \mathbf{F})\} \prod_{i=1}^{I} \pi_i^{z_{i1}} (1-\pi_i)^{z_{i2}} \\ \text{subject to} \quad & \boldsymbol{\pi} \in \mathcal{P}_\beta(\Gamma, \mu_{\pi^*}) \\ & \mu_{\pi^*} \leq \bar{\Gamma}/(1+\bar{\Gamma}). \end{aligned} \tag{3.12}$$

Let $\mathcal{U}_\beta(\Gamma, \bar{\Gamma})$ be the set of feasible solutions to (3.12). Let $\boldsymbol{\pi}_{\text{sup},\beta}$ and $\mu_{\text{sup},\beta}$ be the arg max of (3.12), such that $p(\boldsymbol{\pi}_{\text{sup},\beta}, \mu_{\text{sup},\beta})$ is the tail probability at the solution to (3.12). If $\bar{\Gamma} < \Gamma$, let $p_\beta = p(\boldsymbol{\pi}_{\text{sup},\beta}, \mu_{\text{sup},\beta}) + \beta$; otherwise, let $p_\beta = p(\boldsymbol{\pi}_{\text{sup},\beta}, \mu_{\text{sup},\beta})$.

Proposition 2. *Suppose we sample I pairs from an infinite population through the procedure in §2.1, that treatment assignment is strongly ignorable given (X, U), and that (3.7) and (3.8) hold at Γ and $\bar{\Gamma} \leq \Gamma$ respectively. Then, if H_0 is true, for $0 < \beta \leq 0.5$,*

$$\lim_{I \to \infty} \mathbb{P}(p_\beta \leq \alpha \mid H_0) \leq \alpha$$

That is, p_β is an asymptotically valid p-value for an extended sensitivity analysis testing H_0 with parameters $(\Gamma, \bar{\Gamma})$.

Proof. We first prove the result for $\bar{\Gamma} < \Gamma$. The proof is similar to that of Lemma 1 in Berger and Boos (1994), differing primarily in that the nuisance parameters given $\mathcal{F}_I$, $\boldsymbol{\pi}$, are themselves realizations of random variables in the setting of §3.2.1. Suppose the null hypothesis is true, and let μ_0 be the true value for μ_{π^*}. Further, for any set $\mathcal{F}_I$ let $\boldsymbol{\pi}_0$ be the true value of $\boldsymbol{\pi}$. and let $p(\boldsymbol{\pi}_0, \mu_0)$ be the value of (3.11) evaluated at $\boldsymbol{\pi}_0$ and μ_0.

$$\begin{aligned}
\mathbb{P}(p_\beta \leq \alpha) &= \mathbb{E}[\mathbb{P}(p_\beta \leq \alpha, \bar{\pi}_0^* \in \mathcal{C}_\beta(\Gamma, \mu_0) \mid \mathcal{F}_I)] + \mathbb{E}[\mathbb{P}(p_\beta \leq \alpha, \bar{\pi}_0^* \notin \mathcal{C}_\beta(\Gamma, \mu_0) \mid \mathcal{F}_I)] \\
&\leq \mathbb{E}[\mathbb{P}(p(\boldsymbol{\pi}_0, \mu_0) + \beta \leq \alpha \mid \mathcal{F}_I)] + \mathbb{E}[\mathbb{P}(\bar{\pi}_0^* \notin \mathcal{C}_\beta(\Gamma, \mu_0) \mid \mathcal{F}_I)] \\
&= \mathbb{E}[\mathbb{P}(p(\boldsymbol{\pi}_0, \mu_0) \leq \alpha - \beta \mid \mathcal{F}_I)] + \mathbb{P}(\bar{\Pi}^* \notin \mathcal{C}_\beta(\Gamma, \mu_0))
\end{aligned}$$

The second line follows from $p(\boldsymbol{\pi}_0, \mu_0) \leq \sup_{\boldsymbol{\pi} \in \mathcal{P}_\beta(\Gamma, \mu_0)} p(\boldsymbol{\pi}, \mu_0) \leq p_\beta - \beta$ if $\bar{\pi}_0^* \in \mathcal{C}_\beta(\Gamma, \mu_0)$. By validity of (3.11) at $\boldsymbol{\pi}_0$ given $\mathcal{F}_I$, the first term in the third line is less than or equal to $\alpha - \beta$, while (3.9) illustrates that $\lim_{I \to \infty} \mathbb{P}(\bar{\Pi}^* \notin \mathcal{C}_\beta(\Gamma, \mu_0)) \leq \beta$ for $0 < \beta \leq 0.5$, proving the result for $\bar{\Gamma} < \Gamma$.

If $\bar{\Gamma} = \Gamma$, a solution $\boldsymbol{\pi} \in \mathcal{U}(\Gamma, \Gamma)$ is $\pi_i = \Gamma/(1+\Gamma)$ if $(q_{i1} > q_{i2})$ and $\pi_i = 1/(1+\Gamma)$ otherwise, which recovers the sensitivity analysis of §2.3. Call this solution $\boldsymbol{\pi}_\Gamma$. By arguments in Rosenbaum (2002c, Chapter 4), this solution yields a tight upper bound for the probability in (3.11) under the constraint that $\pi_i^* \leq \Gamma/(1+\Gamma)$. Hence, $p(\boldsymbol{\pi}_{\text{sup},\beta}, \mu_{\text{sup},\beta}) = p(\boldsymbol{\pi}_\Gamma, \Gamma/(1+\Gamma))$ for any β. At $\bar{\Gamma} = \Gamma$, we simply employ the conventional sensitivity analysis which produces valid p-values without an additive increase by β. □

Prior to conducting an extended sensitivity analysis, the practitioner needs to choose a value for β. A compromise must be made, as β acts as a lower bound on the p-value reported by the extended sensitivity analysis but larger values of β correspond to tighter constraints on $\bar{\pi}^*$. Accordingly, we recommend that β be chosen to be smaller than the precision with which p-values are typically reported, but not by much. This recommendation is similar to the guidance given in Berger and Boos (1994).

p_β yields an asymptotically valid p-value for an extended sensitivity analysis with parameters $(\Gamma, \bar{\Gamma})$ because the uncertainty set $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$ defined in (3.10) utilizes the Central Limit Theorem. As our random variables Π_i^* are bounded, we are entitled to certain distribution-free uncertainty sets based on concentration inequalities which have the desired coverage for all sample sizes I; see Appendix 3.8.1 for two approaches using Hoeffding's inequality and Bennett's inequality. These sets, used in place of $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$ when constructing $\mathcal{P}_\beta(\Gamma, \mu_{\pi^*})$, would provide valid p-values for the extended sensitivity analysis through the solution of (3.12) for all values of I. Unfortunately, exact computation of p_β through (3.12) is itself generally intractable, with the additional constraints imposed on the value of $\bar{\pi}$ destroying the properties of the optimization problem solved by the conventional sensitivity analysis which facilitate an exact solution. In §3.4, we provide an implementation of our sensitivity analysis valid in large samples by approximating (3.11) with an appropriate normal distribution, justified under mild conditions. As we employ a normal approximation through our implementation, already implying a large-sample regime, we proceed illustrating the method using the asymptotically valid uncertainty set $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$.

3.3.3. On extended sensitivity analyses for observed study populations

Under the superpopulation model described in §3.2.1, Π_i^* is itself a random variable with expectation $\mathbb{E}[\bar{\Pi}^*]$. In randomized experiments and observational studies, the assumption that the individuals in the study arose as a sample from some larger target population is often specious. Such an assumption is not required for inferential statements, as the act of random assignment to intervention itself can form the basis for probabilistic statements and hypothesis tests, endowing randomized experiments with what Fisher referred to as a "reasoned basis for inference" (Fisher, 1935). Rosenbaum (1999) further argues that the most compelling observational studies are not those which are representative of a larger population, but rather those arrived upon through an active choice of the conditions of observation, seeking the "rare circumstances in which tangible evidence may be obtained to distinguish treatment effects from the most plausible biases" (Rosenbaum, 1999, p. 259).

As (3.5) indicates through conditioning on the study population, $\mathcal{F}_I$, the classical sensitivity analysis in §3.2.3 yields a null distribution for finite-sample inference whose nuisance parameters are the unknown assignment probabilities $\boldsymbol{\pi}$ for the individuals in the study at hand. The parameter Γ, which originally served to bound the supremum of the random variables Π_i^*, also bounds the supremum of the observed values π_i^*. This yields harmony between inference conducted for the finite study population and inference assuming an infinite population into existence when interest is in the hypothesis H_0. Inference given $\mathcal{F}_I$ is valid on its own, but if a superpopulation model is deemed appropriate, inference given $\mathcal{F}_I$ yields valid unconditional inference within that framework.

The motivation for formulating the extended sensitivity analysis with explicit reference to a superpopulation is that while bounds on the supremum of a random variable bound the random variable's realizations, bounds on the expectation of a random variable do not afford bounds in the sample average. The idealized model is used to formulate probabilistic bounds for the sample average $\bar{\Pi}^*$, which then entitle us to a further bound on the average of the realized vector $\boldsymbol{\pi}^*$. Proposition 2 indicates that the price to be paid for implementing this bound is the addition of an extra β term to the p-value, necessitated by the view of $\boldsymbol{\pi}^*$ as a realization of a random variable. Should a superpopulation model be deemed unreasonable, our model could instead be interpreted as placing a bound on the sample average of the parameters $\boldsymbol{\pi}^*$, $\bar{\pi}^*$, in the particular observational study being analyzed. This interpretation eliminates the need for both the uncertainty set $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$ and the increase in the p-value by β, and an option to consider study population inference is available within our `R` function. In our particular case study we proceed using superpopulation bounds, as in calibrating the sensitivity parameters in one observational study by means of another one must assume comparability of biases in the two studies.

3.3.4. A special case: Binary outcomes

Although exact computation of p_β is generally intractable, in one special but common setting it is not. When the outcomes being studied are binary and $t(\mathbf{Z}, \mathbf{F})$ is chosen to be

McNemar's test statistic, computing p_β exactly under Fisher's sharp null $H_0: R_{Tij} = R_{Cij}$ becomes a straightforward exercise. Recall that McNemar's test statistic counts the number of pairs where the subject under treatment has a positive outcome and the control subject does not; that is, $t(\mathbf{Z}, \mathbf{F}) = \sum_{i=1}^{I}(Z_{i1} - Z_{i2})(R_{Ci1} - R_{Ci2})/2 + 1/2$ when Fisher's sharp null is true. Since pairs that are not discordant in treatment and outcome do not contribute to McNemar's statistic it is natural to distinguish pairs that are discordant in outcome and those that are not. Let the first I_d pairs be the discordant pairs and the last I_c be the concordant pairs so that $I = I_d + I_c$. Furthermore, let the first unit of each discordant pair be the unit with positive outcome, that is $R_{i1} = 1$ for $i = 1, \ldots, I_d$.

For the special case of McNemar's test, let μ_m be the value of $\mu_{\pi^*} \leq \bar{\Gamma}/(1+\bar{\Gamma})$ that maximizes the upper bound of $C_\beta(\Gamma, \mu_{\pi^*})$ and let $\bar{\pi}_m$ be the maximized upper bound. Define $\bar{\pi}_c = 1/2$,

$$\bar{\pi}_d = \min\left\{(I\bar{\pi}_m - I_c\bar{\pi}_c)/I_d, \Gamma/(1+\Gamma)\right\},$$

and $\boldsymbol{\pi}_m = ([\bar{\pi}_d \cdot \mathbf{1}_d, \bar{\pi}_c \cdot \mathbf{1}_c])$, where $\mathbf{1}_k$ is a vector of I_k ones. $(\boldsymbol{\pi}_m, \mu_m)$ is then a feasible solution to (3.12) that is designed to put as much bias on the discordant pairs as is allowed by the constraints of the optimization problem. Furthermore, since the concordant pairs do not contribute to the test statistic we have that $p(\boldsymbol{\pi}_m, \mu_m) = \mathbb{P}(B(I_d, \bar{\pi}_d) \geq t(\mathbf{Z}, \mathbf{F}))$, where $B(I_d, \bar{\pi}_d)$ is a Binomial random variable with success probability $\bar{\pi}_d$ and I_d trials. Now, let $p_\beta = p(\boldsymbol{\pi}_m, \mu_m) + \beta$ when $\bar{\Gamma} < \Gamma$ and let $p_\beta = p(\boldsymbol{\pi}_\Gamma, \Gamma/(1+\Gamma))$ otherwise. In the following proposition we show that, in this special setting, an exact solution to (3.12) simply requires computing this Binomial tail probability.

Proposition 3. *Consider a test of $H_0: R_{Tij} = R_{Cij}$ with binary outcomes, and let $t(\mathbf{Z}, \mathbf{F})$ be McNemar's test statistic. Further, let $C_\beta(\Gamma, \mu_{\pi^*})$ be an exact, distribution-free $1-\beta$ uncertainty set. Then under the same conditions as Proposition 2,*

$$\mathbb{P}(p_\beta \leq \alpha \mid H_0) \leq \alpha$$

for any I if $t(\mathbf{Z}, \mathbf{F}) \geq I_d \bar{\pi}_d$. In other words, for any value of I, computing a valid p-value for an extended sensitivity analysis testing H_0 with parameters $(\Gamma, \bar{\Gamma})$ reduces to computing the Binomial tail probability $\mathbb{P}(B(I_d, \bar{\pi}_d) \geq t(\mathbf{Z}, \mathbf{F}))$.

Proof. When $\bar{\Gamma} = \Gamma$, the proof follows immediately from the proof of this case in Proposition 2. Hence, we restrict our attention to the case when $\bar{\Gamma} < \Gamma$. As noted in §3.3.2, if we replace $C_\beta(\Gamma, \mu_{\pi^*})$ with a distribution-free uncertainty set the optimal solution to (3.12) yields a valid p-value for an extended sensitivity analysis for all values of I. All that remains to be shown is that $(\boldsymbol{\pi}_m, \mu_m)$ is the argmax of (3.12).

Without loss of generality, suppose once again that the first subject of each discordant pair is the unit with a positive outcome, $R_{i1} = 1$ for all $i = 1, \ldots, I_d$. Let $(\boldsymbol{\pi}', \mu')$ be a feasible solution of (3.12) and define $\bar{\pi}'_d$ and $\bar{\pi}'_c$ to be the sample average of the maximal assignment probabilities for the discordant and concordant pairs, respectively. $([\bar{\pi}'_d \cdot \mathbf{1}_d, \bar{\pi}'_c \cdot \mathbf{1}_c], \mu')$ is clearly also a feasible solution. Then, Theorem 1 in Hasegawa and Small (2017) implies that $p([\bar{\pi}'_d \cdot \mathbf{1}_d, \bar{\pi}'_c \cdot \mathbf{1}_c], \mu') \geq p(\boldsymbol{\pi}', \mu')$ when $t(\mathbf{Z}, \mathbf{F}) \geq I_d \cdot \bar{\pi}'_d$. Hence, we need only consider feasible solutions of the form $([\bar{\pi}'_d \cdot \mathbf{1}_d, \bar{\pi}'_c \cdot \mathbf{1}_c], \mu')$. An elementary fact about Binomial random variables is that $B(I_d, p_1)$ stochastically dominates $B(I_d, p_2)$ when $p_1 \geq p_2$. By construction, $(\boldsymbol{\pi}_m, \mu_m)$ yields a feasible solution such that $\bar{\pi}_d \geq \bar{\pi}'_d$ for all feasible solutions of the form $([\bar{\pi}'_d \cdot \mathbf{1}_d, \bar{\pi}'_c \cdot \mathbf{1}_c], \mu')$. Consequently, $p(\boldsymbol{\pi}_m, \mu_m) \geq p([\bar{\pi}'_d \cdot \mathbf{1}_d, \bar{\pi}'_c \cdot \mathbf{1}_c], \mu') \geq p(\boldsymbol{\pi}', \mu')$ for all feasible solutions $(\boldsymbol{\pi}', \mu')$ which proves the result for $\bar{\Gamma} < \Gamma$.

□

For McNemar's test, the extended sensitivity analysis exhibits an interesting behavior when $\bar{\pi}_d = \Gamma/(1 + \Gamma)$: the procedure returns a p-value equal to the p-value returned by the conventional sensitivity analysis at Γ *plus* the extra β term. We still pay the cost of specifying a bound on $\mathbb{E}[\Pi_i^*]$ but do not receive the benefit of a tighter constraint on the realization of $\boldsymbol{\pi}^*$ for discordant pairs. What, exactly, explains this phenomenon? A plausible scenario that may give rise to this behavior is when $I_c >> I_d$, i.e. there are many concordant

pairs in the sample of I pairs. In throwing out concordant pairs when using McNemar's statistic, the uncertainty set for $\bar{\Pi}^*$, the average of Π_i^* over all pairs, tells us relatively little about the realized average $\bar{\pi}_d^*$ over discordant pairs, reflecting the cost of bounding the marginal expectation $\mathbb{E}[\Pi_i^*]$ instead of the conditional expectation $\mathbb{E}[\Pi_i^* \mid \mathbf{R}_{Ti}, \mathbf{R}_{Ci}]$.

Although this behavior indicates that the extended sensitivity analysis is, in some sense, suboptimal compared to the conventional sensitivity analysis when $I_c >> I_d$, the practical implications are mostly negligible as β is chosen to be smaller than the precision with which p-values are generally reported. Furthermore, given a choice of Γ and conditional on (I_d, I_c), we can a priori determine the value of $\bar{\Gamma}$ above which the conventional analysis is superior to the extended analysis. Because (I_d, I_c) are known conditional on $\mathcal{F}_I$, we are not at risk of using the data twice – once to choose the best test and once to perform that test. Consequently, the resulting sensitivity analyses will still have the appropriate level.

3.4. Implementation through quadratic programming

The test statistics described in §3.2.3 can be represented as the sum of I independent random variables, $\mathbf{Z}^T\mathbf{q} = \sum_{i=1}^{I} T_i$, where $T_i = (q_{i1} + q_{i2})/2 + (Z_{i1} - Z_{i2})(q_{i1} - q_{i2})/2$. This suggests that, under mild regularity conditions, a central limit theorem would be applicable to the distribution of $\mathbf{Z}^T\mathbf{q}$ for any value of $\boldsymbol{\pi}$ in (3.11) for almost every sample path $\mathcal{F}_I$. One sufficient condition proposed in the special central limit theorem of Hájek et al. (1999, §6.1.2) is that, almost surely,

$$\frac{\sum_{i=1}^{I}(q_{i1} - q_{i2})^2}{\max\limits_{1 \le i \le I}(q_{i1} - q_{i2})^2} \to \infty,$$

which requires that no one term $(q_{i1} - q_{i2})^2$ dominates the sum as the number of pairs increases. (An aside: the central limit theorem in Hájek et al. (1999, §6.1.2) as originally stated applies to sums of the form $\sum_{i=1}^{I} a_i X_i$ where X_i are iid random variables; however, the proof can readily be extended to settings where $I\sigma^2 \le \sum_{i=1}^{I} \mathrm{var}(X_i) \le Ic\sigma^2$ for $c > 1$ while dropping the requirement of identical distribution, which encompasses the setting of

our extended sensitivity analysis). Under a normal approximation, the problem of finding the worst-case p-value is equivalent to finding the worst-case deviate.

Recall that a sensitivity analysis is typically conducted only if the null hypothesis is rejected under the assumption of no unmeasured confounding ($\Gamma = \bar{\Gamma} = 1$), and then proceeds by iteratively increasing the sensitivity parameters until the test fails to reject. Having proceeded to sensitivity analysis only after rejecting the null under no unmeasured confounding, even with one-sided alternatives we can safely consider rejection or failure to reject for sequentially larger values of Γ and $\bar{\Gamma}$ based on the minimal squared deviate, an objective function which is preferred for computational reasons alluded to below. Recalling that under (3.11) we condition on $\mathcal{F}_I$ and hence treat the vector $\mathbf{q}$ as fixed, minimizing the squared deviate can be expressed as an optimization problem over the unknown probabilities $\boldsymbol{\pi}$ as

$$\min_{\boldsymbol{\pi} \in \mathcal{U}_\beta(\Gamma, \bar{\Gamma})} \frac{(t - \mathbb{E}_{\boldsymbol{\pi}}[\mathbf{Z}^T\mathbf{q} \mid \mathcal{F}_I])^2}{\text{var}_{\boldsymbol{\pi}}(\mathbf{Z}^T\mathbf{q} \mid \mathcal{F}_I)}, \tag{3.13}$$

where t is the observed value of the statistic $t(\mathbf{Z}, \mathbf{F})$, and the expectation and variance are for the test statistic $t(\mathbf{Z}, \mathbf{F})$ under the randomization distribution (3.11) for a given vector $\boldsymbol{\pi}$. Under a normal approximation for $t(\mathbf{Z}, \mathbf{F})$, the squared deviate follows a χ^2_1 distribution. By the argument of the previous section, we then reject the null at level α if (3.13) is greater than or equal to $G^{-1}(1-2(\alpha-\beta))$ for one-sided alternatives or $G^{-1}(1-(\alpha-\beta))$ for two-sided alternatives, where $G^{-1}(p)$ is the p quantile of a χ^2_1 distribution.

The expectation and variance of the contribution of T_i can be expressed as a function of the unknown vector $\boldsymbol{\pi}$ as

$$\mathbb{E}_{\boldsymbol{\pi}}[T_i \mid \mathcal{F}_I] = \mathbf{q}_i^T \boldsymbol{\pi}_i \tag{3.14}$$

$$\begin{aligned} \text{var}_{\boldsymbol{\pi}}(T_i \mid \mathcal{F}_I) &= \pi_i(1 - \pi_i)(q_{i1} - q_{i2})^2 \\ &= (\mathbf{q}_i^2)^T \boldsymbol{\pi}_i - (\mathbf{q}_i^T \boldsymbol{\pi}_i)^2 \end{aligned} \tag{3.15}$$

where $\boldsymbol{\pi}_i$ and $\mathbf{q}_i$ are vectors of length two with elements $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2})$ and $\mathbf{q}_i = (q_{i1}, q_{i2})$,

respectively. Suppose without loss of generality that we are considering a one-sided, greater than alternative and that we rejected the null at $(\Gamma, \bar{\Gamma}) = (1, 1)$, which implies that $t \geq (2I)^{-1} \sum_{i=1}^{I} \sum_{j=1}^{2} q_{ij}$ (i.e. that the observed value of t exceeded its null expectation). Sort each vector $\mathbf{q}_i$ in descending order such that $q_{i1} \geq q_{i2}$. Then, $\text{var}_{\boldsymbol{\pi}}(T_i \mid \mathcal{F}_I) = \text{var}_{\boldsymbol{\pi}^*}(T_i \mid \mathcal{F}_I)$ from (3.15), while from (3.14) $\mathbb{E}_{\boldsymbol{\pi}}[T_i \mid \mathcal{F}_I] \leq \mathbb{E}_{\boldsymbol{\pi}^*}[T_i \mid \mathcal{F}_I] = \mathbf{q}_i^T \boldsymbol{\pi}_i^*$ and $(q_{i1} + q_{i2})/2 \leq \mathbb{E}_{\boldsymbol{\pi}^*}[T_i \mid \mathcal{F}_I]$. Hence, any feasible solution $\boldsymbol{\pi}'$ to (3.13) has an objective value that is no smaller than that of $(\boldsymbol{\pi}^*)'$, as the variance will be the same while, recalling the iterative nature of a sensitivity analysis, the distance $(t - \mathbb{E}_{(\boldsymbol{\pi}^*)'}[\mathbf{Z}^T \mathbf{q}' \mid \mathcal{F}_I])^2$ will be smaller than $(t - \mathbb{E}_{\boldsymbol{\pi}'}[\mathbf{Z}^T \mathbf{q}' \mid \mathcal{F}_I])^2$. Maintaining this ordering of the vectors $\mathbf{q}_i$, we can express our optimization problem as a function of the maximal probabilities $\boldsymbol{\pi}_i^*$.

For any candidate $\boldsymbol{\pi}^*$, we reject under a normal approximation with a one-sided, greater than alternative at level $\alpha - \beta$ if the corresponding squared deviate exceeds its critical value, $G^{-1}(1-2(\alpha-\beta))$ i.e. if $\zeta(\boldsymbol{\pi}^*, \alpha - \beta) = (t - \mathbb{E}_{\boldsymbol{\pi}^*}[\mathbf{Z}^T \mathbf{q} \mid \mathcal{F}_I])^2 - G^{-1}(1-2(\alpha-\beta))\text{var}_{\boldsymbol{\pi}^*}(\mathbf{Z}^T \mathbf{q} \mid \mathcal{F}_I) \geq 0$. We write $\zeta(\boldsymbol{\pi}^*, \alpha - \beta)$ explicitly as a function of $\boldsymbol{\pi}^*$ as

$$\zeta(\boldsymbol{\pi}^*, \alpha - \beta) = (t - \mathbf{q}^T \boldsymbol{\pi}^*)^2 - G^{-1}(1 - 2(\alpha - \beta)) \sum_{i=1}^{I} \left((\mathbf{q}_i^2)^T \boldsymbol{\pi}_i^* - (\mathbf{q}_i^T \boldsymbol{\pi}_i^*)^2 \right)$$

If we find that $\zeta(\boldsymbol{\pi}^*, \alpha - \beta) \geq 0$ for all feasible $\boldsymbol{\pi}^* \in \mathcal{U}_\beta(\Gamma, \bar{\Gamma})$, we can reject the null while asymptotically controlling the size of the extended sensitivity analysis with parameters $(\Gamma, \bar{\Gamma})$ at α. The function $\zeta(\boldsymbol{\pi}^*, \alpha - \beta)$ is convex and quadratic in $\boldsymbol{\pi}^*$. Meanwhile, we explicitly write the constraints determining membership in $\mathcal{U}_\beta(\Gamma, \bar{\Gamma})$ as

$$1/2 \leq \pi_i^* \leq \Gamma/(1+\Gamma), \quad 1 \leq i \leq I \tag{3.16}$$

$$I^{-1} \sum_{i=1}^{I} \pi_i^* \leq \mu_{\pi^*} + I^{-1/2} \Phi^{-1}(1-\beta) \left\{ (\Gamma/(1+\Gamma) - \mu_{\pi^*}) (\mu_{\pi^*} - 1/2) \right\}^{1/2} \tag{3.17}$$

$$\mu_{\pi^*} \leq \bar{\Gamma}/(1+\bar{\Gamma}). \tag{3.18}$$

For a fixed value of $\mu_{\pi^*} \leq \bar{\Gamma}/(1+\bar{\Gamma})$ the constraints are linear in the unknown maximal

probabilites π_i^*. Hence, for fixed μ_{π^*}, the problem $\min_{\boldsymbol{\pi}^*} \zeta(\pi^*, \alpha - \beta)$ subject to (3.16) and (3.17) can be written as a quadratic program. With a one-sided alternative, an asymptotically level-α extended sensitivity analysis with parameters $(\bar{\Gamma}, \Gamma)$ simply requires checking whether the solution to that quadratic program is greater than or equal to zero, rejecting the null if so and failing to reject otherwise. For a two-sided alternative, simply replace $\zeta(\pi^*, \alpha - \beta)$ with $\zeta(\pi^*, (\alpha - \beta)/2)$ to control the level of the procedure at α. See Rosenbaum (1992) and Fogarty and Small (2016) for similar formulations of sensitivity analyses as convex programs.

A minor complication is that for small values of I or for small values for β, the right-hand side of (3.17) need not be monotone increasing in μ_{π^*} if $2\bar{\Gamma}/(1+\bar{\Gamma}) \geq \Gamma/(1+\Gamma) + 1/2$, as decreasing μ_{π^*} may lead to an increase in the component dependent on the variance bound which exceeds the corresponding decrease in the additive term μ_{π^*}. To remedy this, one can simply find the value for μ_{π^*} over the range $[(\Gamma/(1+\Gamma) + 1/2)/2, \bar{\Gamma}/(1+\bar{\Gamma})]$ which maximizes the right-hand side of (3.17) through a bisection algorithm, and then proceed with the quadratic program using this single value. If $2\bar{\Gamma}/(1+\bar{\Gamma}) < \Gamma/(1+\Gamma)+1/2$, the right-hand side of (3.17) is, subject to (3.18), maximized at $\mu_{\pi^*} = \bar{\Gamma}/(1+\bar{\Gamma})$, so one can proceed by replacing μ_{π^*} with $\bar{\Gamma}/(1+\bar{\Gamma})$ and solving the required quadratic program. Importantly, the method only requires solving a single quadratic program. Quadratic programs can be solved by many free and commercially available solvers; we provide code implementing our method using the `R` package for the solver `Gurobi`, which is free for academic use, at the author's website `http://www.raidenhasegawa.com`. We also provide options to replace the constraint (3.17), justified by the Central Limit Theorem, with bounds described in Appendix 3.8.1 which are valid for any I through distribution-free concentration inequalities.

3.5. Simulations

3.5.1. Type I error control

In the following simulations, we demonstrate that the extended sensitivity analysis introduced in §3.3 has the correct level. We consider two important cases: (1) when no unmeasured bias is present and (2) when the there is unmeasured bias but the sensitivity analysis is conducted at the true values of Γ and $\bar{\Gamma}$. In both settings we test Fisher's sharp null that $\tau = 0$ using the difference in means test with desired Type I error control at $\alpha = 0.05$. We set $\beta = \alpha/10 = 0.005$ for conducting the extended sensitivity analysis. The following treatment model, outcome model, and simulation settings were used to conduct the Type I error control simulations:

1. **Treatment model:** $\Pi_i^* = 1/2$ with probability $p = 2(\Gamma - \bar{\Gamma})/\{(\Gamma - 1)(\bar{\Gamma} + 1)\}$ and $\Pi_i^* = \Gamma/(1 + \Gamma)$ with probability $1 - p$.

2. **Outcome model:**

   - *unbiased:* $Y_i = \tau \cdot (Z_{i1} - Z_{i2}) + \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$,

   - *biased:* $Y_i = \tau \cdot (Z_{i1} - Z_{i2}) + \{2 \cdot \chi(\pi_i > 1 - \pi_i) - 1\} \cdot |\epsilon_i|$ where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, 1)$.

3. **Sensitivity parameters:**

   - $\Gamma \in \{1, 1.1, 1.25, 1.5, 2\}$,

   - $\bar{\Gamma} \in \{1, 1.05, 1.1, 1.15, 1.2, 1.25, 1.3, 1.35,$
     $1.4, 1.45, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0\}$,

   - $\bar{\Gamma} \leq \Gamma$.

4. **Study and simulation size:** $I = 100$ pairs, $N_{sim} = 5000$ simulations.

In the biased setting, the unit with higher potential outcome under control has higher

probability of receiving treatment. When $\Gamma = \bar{\Gamma} = 1$ we use the convention that $p = 0/0 = 0$. The value of $p = \mathbb{P}(\Pi_i^* = 1/2)$ was chosen so that the population treatment model satisfies $\mathbb{E}[\bar{\Pi}^*] = \bar{\Gamma}/(1+\bar{\Gamma})$. The results of the simulation study for the biased and unbiased settings are shown in Table 6 and the Table 9 in §3.8.5 of the Appendix, respectively. The extended sensitivity procedure correctly controls the Type I error rate for all pairs of sensitivity parameters $(\Gamma, \bar{\Gamma})$ tested. The first row of each table, where $\bar{\Gamma} = 1$, corresponds to tests under the absence of unmeasured confounding. The pairs where $\Gamma = \bar{\Gamma}$ correspond to the conventional worst-case sensitivity analysis. Under the unbiased treatment model, the extended sensitivity analysis is typically more conservative as we increase Γ or $\bar{\Gamma}$. In the biased setting, we observe the same pattern as we vary Γ, but as $\bar{\Gamma}$ approaches Γ, the level of the extended sensitivity analysis does not decrease monotonically. In fact, at a certain value of $\bar{\Gamma}$, the extended sensitivity analysis becomes less conservative as we approach Γ. In short, the solution $\boldsymbol{\pi}_{sup,\beta}$ to the optimization problem in (3.12) tends to more closely approximate the true allocation $\boldsymbol{\pi}_0$ when $\bar{\Gamma}$ is close to either 1 or Γ in the biased setting. When $\bar{\Gamma}$ is close to 1, the feasible set of $\boldsymbol{\pi}$'s is closely bounded around $\boldsymbol{\pi}_0 \approx \mathbf{1} \cdot 1/2$. When $\bar{\Gamma}$ is close to Γ the true allocation is $\boldsymbol{\pi}_0 \approx \boldsymbol{\pi}_\Gamma$ and the extended sensitivity analysis behaves like the conventional sensitivity analysis, where $\boldsymbol{\pi}_{\text{sup},\beta} = \boldsymbol{\pi}_\Gamma$ yields a tight upper bound on the probability in (3.11). In between these edge cases, when the feasible set of $\boldsymbol{\pi}$ is relatively large and the trade-off between maximizing expectation and variance is more nuanced, (3.12) may produce solutions $\boldsymbol{\pi}_{\text{sup},\beta}$ that yield appreciably more conservative inference than if had we known the true $\boldsymbol{\pi}_0$.

3.5.2. The power of an extended sensitivity analysis

The power of a sensitivity analysis quantifies the ability of an observational study design to distinguish treatment effects from unmeasured bias. Formally, it reports for a given study design the probability of rejecting a false null hypothesis for a chosen level α and sensitivity parameter Γ under 'favorable' conditions, defined in Rosenbaum (2010, Chapter 14), as the presence of a treatment effect that causes meaningful effects and absence of

			Γ		
$\bar{\Gamma}$	1	1.1	1.25	1.5	2
1	0.047	0.047	0.045	0.046	0.044
1.05		0.022	0.011	0.007	0.005
1.1		0.032	0.010	0.004	0.003
1.15			0.012	0.002	0.002
1.2			0.017	0.004	0.001
1.25			0.025	0.004	0.001
1.3				0.006	0.000
1.35				0.009	0.001
1.4				0.011	0.001
1.45				0.014	0.001
1.5				0.025	0.001
1.6					0.003
1.7					0.004
1.8					0.006
1.9					0.011
2					0.021

Table 6: Rejection probability of the true null hypothesis, $H_0 : \tau = 0$, under the *biased* setting with target Type I error control at $\alpha = 0.05$. The Monte Carlo standard error of these probability estimates is bounded above by $\sqrt{0.05 \times 0.95/5000} \approx 0.003$ if the true Type I error rate is 0.05.

unmeasured biases. The investigator cannot determine from observable data alone whether or not such favorable conditions hold. An attractive study design would be highly insensitive to unmeasured confounding if she was lucky enough to find herself in this favorable setting. The power of an extended sensitivity analysis extends this formalism to the triplet $(\alpha, \Gamma, \bar{\Gamma})$. Power simulations for $\alpha = 0.05$ and several pairs of $(\Gamma, \bar{\Gamma})$ are reported in Table 7 and Table 10 in §3.8.5 of the Appendix for $\tau = 0.5$ and $\tau = 0.25$, respectively. Other than the presence of a 'meaningful' treatment effect τ, the simulation settings are identical to the unbiased setting in §3.5.1.

Unsurprisingly, the power of the extended sensitivity analysis decreases as $\bar{\Gamma}$ approaches Γ. If the investigator has reason to believe that unmeasured confounding is heterogeneous and that extreme pairwise unmeasured confounding is possible but relatively rare, the conventional sensitivity analysis is likely unduly conservative. Further, the extended sensitivity analysis allows the investigator to compare the power of competing study designs under

different assumptions about the maximal and expected degree of unmeasured confounding.

	$\boldsymbol{\Gamma}$				
$\bar{\boldsymbol{\Gamma}}$	1	1.1	1.25	1.5	2
1	0.998	0.999	0.998	0.999	0.999
1.05		0.994	0.990	0.984	0.978
1.1		0.996	0.984	0.965	0.941
1.15			0.977	0.947	0.896
1.2			0.978	0.928	0.833
1.25			0.979	0.907	0.759
1.3				0.890	0.719
1.35				0.884	0.664
1.4				0.879	0.626
1.45				0.874	0.578
1.5				0.882	0.541
1.6					0.505
1.7					0.478
1.8					0.463
1.9					0.472
2					0.486

Table 7: Rejection probability of the false null hypothesis, $H_0 : \tau = 0$, under the *unbiased* setting with true alternative hypothesis $H_1 : \tau = 0.5$. The Monte Carlo standard error of these probability estimates is bounded above by $\sqrt{0.5 \times 0.5/5000} \approx 0.007$.

3.6. Extended sensitivity analysis for returns to schooling

3.6.1. A model for returns to schooling

How does going to college affect job earnings? The question and the implications of the many putative answers are important to education policy experts and parents alike. It has been empirically demonstrated that log earnings are nearly a linear function of schooling (see, for instance, Card and Krueger, 1992). In the idealized paired observational setting introduced in §§3.2.1-3.2.2 where the treatment condition is attending college for at least two years and the control condition is receiving at most a high school diploma, a hypothesized treatment effect $\tau \times 100$ would describe the percentage increase in earnings associated with attending at least two years of college, the minimum number of years to receive an associates degree. Formally, we consider the multiplicative treatment effect hypothesis $H_\tau : R_{Tij} = \tau R_{Cij}$ where (R_{Tij}, R_{Cij}) are potential earnings after attending college or not.

Choosing $t(\mathbf{Z}, \mathbf{F}) = \mathbf{Z}^T\mathbf{q}$ to be the adjusted difference-in-means test comparing log earnings, q_{ij} would take the form $q_{ij} = (\log R_{Tij} - \log R_{Cij'}) - \log(\tau)$ and $q_{ij'} = -q_{ij}$ under H_τ.

Let $X = [X_f, X_s]$ where X_f and X_s are familial and subject level covariates. In an idealized sibling comparison design, the strong ignorability condition in (3.1) would hold with respect to X_f; that is, if for all x_f,

$$(R_T, R_C) \perp\!\!\!\perp Z \mid X_f, \;\; 0 < \mathbb{P}(Z = 1 \mid X_f = x_f) < 1. \tag{3.19}$$

If X_s does not affect treatment assignment but does predict potential outcomes, this sibling version of strong ignorability will still hold. For example, in the sibling pairs from the WLS data that we consider in the following section, the age at which income is measured (AGE) is different between siblings. If $X_s = AGE$, then it is conceivable that X_s does not affect whether a sibling went to college or not. This would not be the case for people who went to college later in life or whose family characteristics may have changed over time, in which case AGE would be a proxy for those changes. Regardless, model-agnostic adjustment for X_s and X_f can improve the power of the resulting sensitivity analysis (Rosenbaum, 2002b). For example, we can use simple linear regression to adjust for X by replacing $\mathbf{q}$ with $(I - H_{X_s})\mathbf{q}$ where H_{X_s} is the orthogonal projection onto X_s without an intercept.

3.6.2. Ashenfelter: Conventional versus extended sensitivity analysis

To illustrate the differences between the conventional and extended sensitivity analyses, we return to the twin study of Ashenfelter and Rouse (1998) (AR). AR collected survey data on 680 monozygotic twins (340 pairs) attending the Twinsburg Twins Festival in Twinsburg, Ohio during the summers of 1991, 1992, and 1993. We consider the 40 pairs of twins where one twin attend at least two years of college and the other had no more than a high school education, and where both twins were employed at the time of data collection. Assuming no unmeasured confounding, testing Fisher's sharp null H_0 yields a p-value of ≈ 0.0001. We obtain a 95% confidence interval for $\log(\tau)$ of [0.16,0.43] by inverting H_τ for $\tau \in \mathbb{R}_+$ at

$\alpha = 0.05$ with a two-sided alternative. Exponentiating the endpoints, attending at least two years of college versus receiving at most a high school diploma increased wages by between 17% and 53% with 95% confidence.

Being a retrospective study neither baseline IQ nor any other intelligence scores were collected, and a critical reader may point to the possible presence of ability bias as a basis to call the conclusions of the study into question. Conducting a sensitivity analysis produces a quantitative rejoinder to this type of criticism in the form of a *sensitivity value* Γ^* for the conventional analysis and a *sensitivity curve* $(\Gamma^*, \bar{\Gamma}^*)$ for the extended analysis. The sensitivity value is the largest bound on the maximal bias such that the qualitative conclusions of the study do not change (i.e., such that we reject H_0). The sensitivity curve is the two-dimensional analog of the sensitivity value and can be seen as the threshold between the gray region (reject H_0) and the white region (retain H_0) in Figure 3. At the limits of the sensitivity curve, we recover two separate single-parameter sensitivity analyses. The sensitivity value returned by the conventional analysis corresponds to the point where the sensitivity curve intersects the $y = x$ line ($\Gamma^* \approx 2.36$). The limit of the sensitivity curve as $\Gamma \to \infty$ is the sensitivity value of a single-parameter sensitivity analysis that bounds the typical bias ($\bar{\Gamma} \approx 1.22$).

3.6.3. Ability Bias: Cross-study sensitivity analysis calibration

Without context, the sensitivity curve and values from the Ashenfelter analysis may be difficult to interpret. In response to the critic of the "equal abilities" hypothesis for twins, we would ideally like to report whether or not the Ashenfelter study is sensitive to plausible patterns of ability bias. One strategy for addressing this is to estimate the bias due to ability from a *calibration study* that has a comparable design and information on baseline ability such as IQ. We can then *calibrate* the sensitivity analysis to these estimates of Γ and $\bar{\Gamma}$. To implement this *cross-study calibration*, we modify the procedure established in Hsu and Small (2013) to calibrate sensitivity parameters to observed covariates. In brief, one fits ostensible treatment and outcome models – for instance, via linear and logistic regression

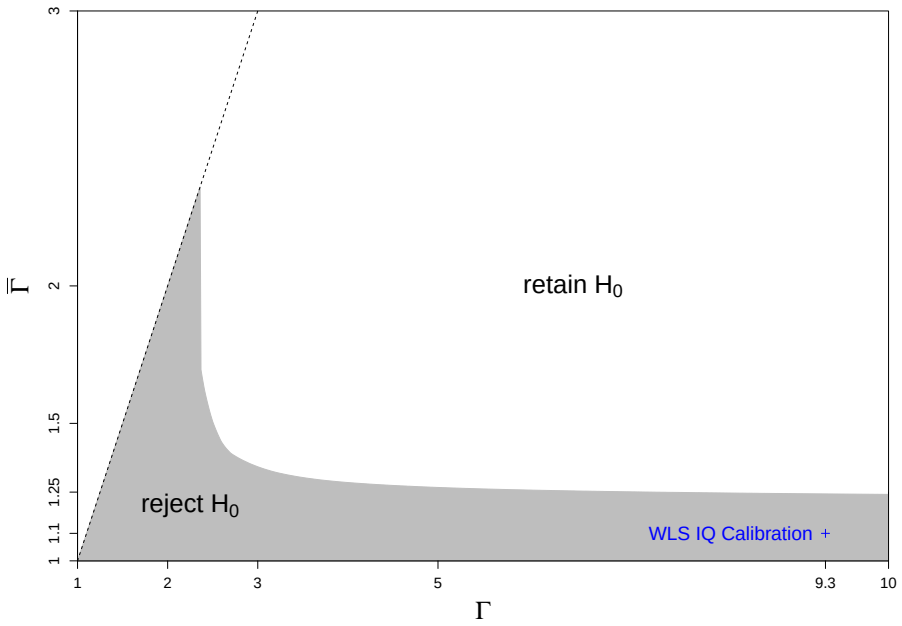


Figure 3: Extended sensitivity curve from the AR study calibrated to the estimates of ability bias from the WLS study (cross). The gray region indicates the sensitivty parameter pairs $(\Gamma, \bar{\Gamma})$ for which H_0 can still be rejected. The point where the sensitivity curve intersects the $y = x$ line corresponds to the sensitivity value returned by conventional sensitivity analysis ($\Gamma^* \approx 2.36$). The limit of the curve as $\Gamma \to \infty$ corresponds to the sensitivity value returned by the single-parameter sensitivity analysis that bounds the typical bias ($\bar{\Gamma}^* \approx 1.22$).

– and uses the resulting model fits to estimate $\boldsymbol{\pi}^*$, $\bar{\Gamma}$, and Γ. The details of this step can be found in Appendix 3.8.3. Calibrating the sensitivity analysis to estimates of ability bias provides the context relevant to the critic's concerns.

To assess the robustness of the AR study to ability bias, we use the sibling data from the WLS study introduced in §3.1.2 to design a calibration study. We constructed a set of 171 same-sex, full-sibling pairs that received discordant treatment. We let $Z_{ij} = 0$ if sibling j in pair i received 12 or fewer years of education and $Z_{ij} = 1$ if he or she received 14 or more years of education (at least two years of college). Log income for the previous year was collected for WLS participants and their siblings in 1975 and 1977, respectively. To more closely approximate the superpopulation from which the AR twins came, we only consider siblings where both had non-zero income at the time of collection (i.e. were employed). As

outlined in the previous section, we let $X_s = AGE$ and use regression to adjust $\mathbf{q}$ for the age at which income was collected. This calibration analysis is stylized to some extent to avoid obscuring the primary contribution of our method. Many other subject-level covariates are available for adjustment via regression. A detailed analysis including treatment modification with respect to gender and more thorough covariate adjustment would not preclude the use nor usefulness of our method.

Using the 171 WLS sibling pairs, we estimate that $\Gamma \approx 9.3$ and $\bar{\Gamma} \approx 1.1$, summarizing the information we have about maximal and typical biases due to IQ disparities. Heterogeneneity of ability bias can explain the considerable difference between these two measures of confounding. The histogram of the estimated $\boldsymbol{\pi}^*$ in Figure 4 indicates that most sibling pairs have modest differences in intelligence in high school but in a few rare cases the disparity in sibling IQ exposes pairs to high levels of bias. Calibrating the conventional sensitivity analysis of AR to the WLS study would suggest that our conclusions are likely not robust to plausible patterns of ability bias since $\Gamma^* < 9.3$. However, calibration of the extended sensitivity analysis suggests otherwise. In Figure 3, the WLS IQ calibration point $(9.3, 1.1)$ is indicated by the blue cross and falls below the sensitivity curve. The single-parameter sensitivity analysis that bounds the typical bias agrees with the extended analysis that the conclusions are robust to plausible patterns of ability bias ($\bar{\Gamma}^* \geq 1.1$). Incorporating information about the heterogeneity of ability bias by bounding both the maximal and typical biases promotes a less pessimistic assessment of an observational study's robustness to unmeasured confounding. When information on the heterogeneity of potential confounders is available, as in the above cross-study calibration analysis, the extended sensitivity analysis provides a richer picture of the study's robustness to hidden bias.

3.6.4. Sensitivity intervals: Interval estimates with hidden bias

For a fixed bound on the worst-case bias, incorporating heterogeneous bias through the extended sensitivity can also produce narrower *sensitivity intervals* than those attained through the conventional analysis. Representing a natural extension of confidence intervals

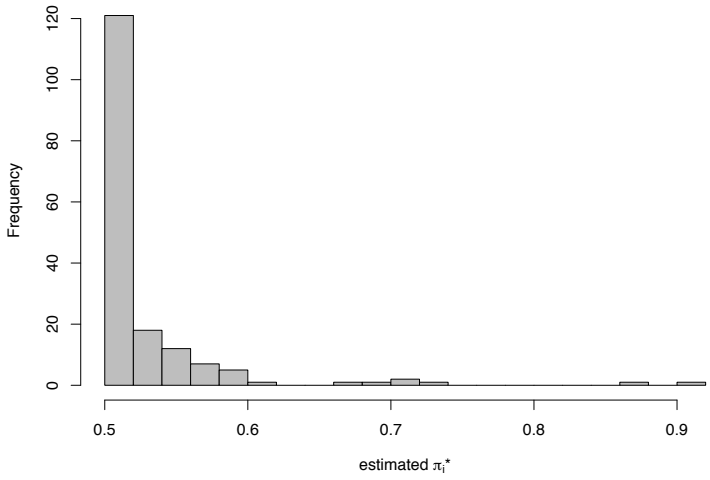


Figure 4: Histogram of $\boldsymbol{\pi}^*$ estimated for 171 same-sex, full-sibling pairs from the WLS study.

to inference in the presence of unmeasured confounding, a $100(1-\alpha)\%$ sensitivity interval is constructed by inverting a level-α extended sensitivity analysis with a two-sided alternative at a given pair of values $(\Gamma, \bar{\Gamma})$. Explicitly, let $p_\beta(\Gamma, \bar{\Gamma}, \tau)$ be the two-sided p-value bound returned by the extended sensitivity analysis in (3.12) for particular values of Γ and $\bar{\Gamma}$. Then, a $100(1-\alpha)\%$ sensitivity interval can be written as $\mathcal{I}(\{\tau : p_\beta(\Gamma, \bar{\Gamma}, \tau) \leq \alpha\})$, where $\mathcal{I}(A)$ is the smallest interval containing the set A. At $\Gamma = \bar{\Gamma} = 1$, the sensitivity interval is simply the corresponding confidence interval found by inverting H_τ using the randomization p-value given in (3.2) as would be justified in a paired experiment. Setting $\Gamma = \bar{\Gamma} > 1$ returns sensitivity intervals produced through the conventional sensitivity analysis, while setting $\Gamma > \bar{\Gamma} > 1$ employs the extended sensitivity analysis in constructing the sensitivity intervals.

Table 8 illustrates the potential for reduced interval lengths through accommodating heterogeneity in unmeasured confounding. It reports 95% sensitivity intervals for $\log(\tau)$ in the AR study with three pairs of values for Γ and $\bar{\Gamma}$. The first, denoted by $\mathcal{I}_{rand}$, is the 95% sensitivity interval assuming no unmeasured confounding previously reported in §6.2. The

second, $\mathcal{I}_{sup}$, is the 95% sensitivity interval derived by setting $\Gamma = \bar{\Gamma} = 9.3$, the calibrated value of the maximal bias parameter from the WLS study. This is precisely the sensitivity interval that the conventional sensitivity analysis bounding only the worst-case confounding would return. The final interval, $\mathcal{I}_{ext}$, is the 95% sensitivity interval setting $\Gamma = 9.3$, $\bar{\Gamma} = 1.1$ in accord with the calibrated values of the maximal and typical bias from the WLS study. We see that $\mathcal{I}_{ext}$ is more than 80% shorter than $\mathcal{I}_{sup}$. Further, both $\mathcal{I}_{rand}$ and $\mathcal{I}_{ext}$ exclude zero while $\mathcal{I}_{sup}$ does not. The positive finding in the unconfounded setting can be explained away by bias calibrated to the WLS study using the conventional sensitivity model, but not when using the extended sensitivity model. Once again, we see that when it is plausible that the typical bias to which pairs are subject is materially smaller than the worst-case bias, the conventional analysis may be overly pessimistic about how informative the data is.

Interval Type	95% Sensitivity Interval
$\mathcal{I}_{rand}$	[0.16,0.43]
$\mathcal{I}_{sup}$	[-0.88,1.63]
$\mathcal{I}_{ext}$	[0.06,0.53]
$100 \times (1 - \lvert\mathcal{I}_{ext}\rvert / \lvert\mathcal{I}_{sup}\rvert)$	81%

Table 8: 95% sensitivity intervals for $\log(\tau)$ in the AR study constructed by inverting H_τ for different values of Γ and $\bar{\Gamma}$. $\mathcal{I}_{rand}$ is the 95% confidence interval for $\log(\tau)$ in the unconfounded setting, $\Gamma = \bar{\Gamma} = 1$. $\mathcal{I}_{sup}$ and $\mathcal{I}_{ext}$ are 95% sensitivity intervals derived from the conventional sensitivity analysis and the extended sensitivity analysis respectively. These intervals are formed using the sensitivity parameters calibrated from the WLS data, $(\Gamma, \bar{\Gamma}) = (9.3, 1.1)$. The percentage reduction in interval length from accommodating heterogeneous unmeasured confounding, $100 \times (1 - |\mathcal{I}_{ext}|/|\mathcal{I}_{sup}|)$, is reported in the last row.

3.7. Concluding remarks

While convenient for ease of calculation, the low-dimensional sensitivity analysis bounding the supremum may fail to address specific concerns with unmeasured confounding in certain contexts. Rosenbaum and Silber (2009) present an amplification of the conventional sensitivity analysis, where the one-dimensional analysis based on Γ is mapped to a curve of two-dimensional analyses which simultaneously bound the extent to which differences

in unobserved covariates can influence the odds of being treated and the odds of having a higher potential outcome under control by the pair (Λ,Δ). This amplificiation provides an aid to interpretation, allowing the researcher to posit bounds on the extent to which unmeasured confounding can affect treatment decisions and the outcome variable. Rather than amplifying the conventional sensitivity analysis, the extended sensitivity analysis provides the researcher a way to further control the distribution of the unmeasured confounders beyond bounding the supremum. In fact, amplification and extension can be viewed as complementary tools available to the researcher. It is straightforward to employ both: the conventional supremum bound Γ that appears in the extended sensitivity analysis may be amplified yielding yet an even richer analysis, with $\bar{\Gamma}$ bounding the typical probability that the treated individual in a pair has the larger (smaller) potential outcome under control for greater-than (less-than) alternatives.

Framing sensitivity analysis in terms of the typical bias is not a new idea, but has been largely unaddressed in the literature; the idea of expected bias appears briefly in Wang and Krieger (2006) in the context of population-level inference for binary outcomes but is not the focus of the paper. In a particular sense, Cornfield et al. (1959) anticipated the duality of both amplified and extended sensitivity analyses in their seminal work on sensitivity analysis. In their smoking and lung cancer example, the authors considered a hypothetical hormone X which increases the probability of developing lung cancer among those exposed from r_2 to r_1 and due to a positive correlation between exposure to X and smoking, appears in a higher proportion among smokers than non-smokers (i.e $p_1 > p_2$). At once, Cornfield et al. (1959) captures the spirit of an amplified analysis in specifying how X is related to both treatment assignment and outcome and that of an extended analysis by imagining that hormone X is not completely absent among non-smokers and completely present among smokers, leading to exposure to bias that is heterogeneous across subjects within both groups.

The concept of heterogeneous unmeasured confounding appeared naturally, if not intention-

ally, in Cornfield's original example. The extended sensitivity analysis introduced in this paper brings this idea into a modern light and provides the researcher with a way to conduct a sensitivity analysis while bounding both maximal and typical biases in matched pair studies. Using two sibling studies on the returns of schooling to income, we demonstrated that a sensitivity analysis bounding the maximal *and* typical bias is both natural and less susceptible to an overly pessimistic view of the study's robustness to hidden bias. When a researcher believes that most, if not all, pairs are exposed to the worst-case bias, our procedure can recover the conventional analysis by setting $\bar{\Gamma} = \Gamma$. If however, the researcher is worried that some, though few, pairs may be exposed to arbitrarily large biases all is not lost; by letting Γ tend to ∞ the extended sensitivity analysis recovers a single-parameter sensitivity analysis that bounds the typical bias.

3.8. Appendix

3.8.1. Construction of Valid Finite-Sample Uncertainty Sets

We now describe the construction of two $100(1-\alpha)\%$ uncertainty sets for Π^* valid for any number of pairs I. The first is based on Hoeffding's inequality, which implies that the set

$$\mathcal{H}_\beta(\Gamma, \mu_{\pi^*}) = (-\infty, \mu_{\pi^*} + I^{-1/2}\left\{1/2\log(1/\beta)(\Gamma/(1+\Gamma) - 1/2)^2\right\}^{1/2}]$$

satisfies $\mathbb{P}(\bar{\Pi} \in \mathcal{H}_\beta(\Gamma, \mu_{\pi^*})) > 1-\beta$ for all values of I. The second combines Bennett's inequality and the Bhatia-Davis inequality to create the set

$$\begin{aligned}\mathcal{B}_\beta(\Gamma, \mu_{\pi^*}) &= (-\infty, \bar{\mu}_{\pi^*} + b_\beta(\Gamma, \mu_{\pi^*}, I)] \\ b_\beta(\Gamma, \mu_{\pi^*}, I) &= \texttt{SOLVE}\{a : I^{-1}\log(1/\beta)(\Gamma/(1+\Gamma) - 1/2)^2/\nu^2(\Gamma, \mu_{\pi^*}) = \\ &\qquad h\left(a(\Gamma/(1+\Gamma) - 1/2)/\nu^2(\Gamma, \mu_{\pi^*})\right)\},\end{aligned}$$

where $h(x) = (1+x)\log(1+x) - x$. $\mathcal{B}_\beta(\Gamma, \mu_{\pi^*})$. This set also satisfies $\mathbb{P}(\bar{\Pi} \in \mathcal{B}_\beta(\Gamma, \mu_{\pi^*})) > 1-\beta$ for any I if $\mathbb{E}[\bar{\Pi}^*] = \mu_{\pi^*}$.

In practice, the upper bound of the set based on Bennett's inequality is smaller than that based on Hoeffding's inequality when μ_{π^*} is far from $(\Gamma/(1+\Gamma)+1/2)/2$, while the ordering reverses when μ_{π^*} is close to the midpoint. The price paid for this exactness for any I is that the upper bounds for both intervals are larger than those of $\mathcal{C}_\beta(\Gamma, \mu_{\pi^*})$, the asymptotically valid uncertainty set based on the Central Limit Theorem.

As noted in the manuscript, the general reliance of our implementation on asymptotic normality reduces the attractiveness of these finite sample uncertainty sets; however, in the case of McNemar's test with binary data, employing either $\mathcal{H}_\beta$ or $\mathcal{B}_\beta$ yields an extended sensitivity analysis for Fisher's sharp null valid for any sample size. `R` functions to compute these uncertainty set can be found in the file `multipliers.R` at the author's website `http://www.raidenhasegawa.com`.

3.8.2. Constructing the WLS same-sex sibling sample

Of the 10,317 individuals in the WLS sample, 7,928 had a randomly chosen sibling who was surveyed. Of those 7,928 subjects with sibling data, 2,106 had information about sibling status (i.e. full, half or step siblings) of which 2,004 were full siblings. 1,486 of these sibling pairs were same-sex siblings of which 49.3% were men. Of the same-sex sibling pairs, there were 749 (40.6% men) where both had no more than a high school education, 265 (64.9% men) where both had at least two years of college education, and 323 (58.8% men) where one had at most a high school degree and the other had at least two years of college education. Of the same-sex pairs discordant in educational attainment, 171 (74.9% men) had complete IQ data and non-zero reported income. There were 149 (45.0% men) same-sex sibling pairs of the 1,486 for which the treatment and control conditions were not well defined – at least one sibling had only one year of college education.

3.8.3. Calibrating Sensitivity Parameters to Disparities in IQ in the WLS Study

We follow a modified version of the calibration strategy introduced in Hsu and Small (2013) which involves estimating putative treatment and outcome models as a function of (X, U)

under H_0 via maximum likelihood where the likelihood is marginalized over the unknown confounder U. Our modification is as follows: instead of marginalizing over the unobserved covariate we suppose that the only unobserved confounder in the Ashenfelter study is intelligence, which is measured via baseline IQ scores in the WLS study. Consequently, estimating the bias due to IQ disparities using the WLS data permits a cross-study calibration of the Ashenfelter and Rouse sensitivity analysis.

By definition, X_f is controlled automatically between siblings. We make the stylized assumption that $X_s = AGE$. Further, we assume that AGE does not affect treatment assignment. Finally, we assume that intelligence is the only unmeasured confounder in the Ashenfelter and Rouse study (i.e. $U = IQ$). Under these assumptions, a possible model for treatment assignment is

$$\mathbb{P}(Z_{ij} = 1 \mid X_{f,i},\, X_{s,ij},\, U_{ij}) = \frac{\exp(\alpha_{Z,i} + \beta_{Z,IQ} \cdot IQ_{ij})}{1 + \exp(\alpha_{Z,i} + \beta_{Z,IQ} \cdot IQ_{ij})}\,. \tag{3.20}$$

The pair specific intercept $\alpha_{Z,i}$ captures the $X_{f,i}$ effects. We estimate the treatment model using conditional likelihood maximization using the `R` function `clogit` in order to avoid bias arising from the fact that the number of α_i to be estimated grows with the sample size. We consider a Gaussian linear model for the outcome

$$Y_{ij} = \alpha_{Y,i} + \beta_{Y,AGE} \cdot AGE_{ij} + \beta_{Y,IQ} \cdot IQ_{ij} + \epsilon_{ij} \quad \text{such that } \epsilon_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)\,. \tag{3.21}$$

We estimate the treatment assignment and outcome models using the 171 discordant sibling pairs that we analyze from the WLS study in the paper.

In the Ashenfelter and Rouse twins study, AGE is controlled within twin pairs so we are interested in calibrating the sensitivity parameters to the estimated bias due to IQ disparities alone. Following Hsu and Small (2013) we estimate that, controlling for age and assuming that IQ is the only confounding factor, the probability that the sibling that went

to college reported a higher income in pair i to be

$$\pi_i(\mathbf{IQ}) = \frac{\exp\{\hat{\beta}_{Z,IQ}(IQ_{i1} - IQ_{i2})\}\exp\{(\hat{\beta}_{Y,IQ}/\hat{\sigma}^2)(Y_{i(2)} - Y_{i(1)})(IQ_{i1} - IQ_{i2})\} + 1}{[1 + \exp\{\hat{\beta}_{Z,IQ}(IQ_{i1} - IQ_{i2})\}][1 + \exp\{(\hat{\beta}_{Y,IQ}/\hat{\sigma}^2)(Y_{i(2)} - Y_{i(1)})(IQ_{i1} - IQ_{i2})\}]}$$

where $Y_{i(1)} = \min\{Y_{i1}, Y_{i2}\}$ and $Y_{i(2)} = \max\{Y_{i1}, Y_{i2}\}$. Define $\boldsymbol{\pi}(\mathbf{IQ})$ to be the 171×1 vector of $\pi_i(\mathbf{IQ})$. Letting $\boldsymbol{\pi}^*(\mathbf{IQ}) = \boldsymbol{\pi}(\mathbf{IQ})$ when $\hat{\beta}_{Z,IQ}\hat{\beta}_{Y,IQ} \geq 0$ and $1 - \boldsymbol{\pi}(\mathbf{IQ})$ otherwise, one reasonable set of estimates for $(\Gamma, \bar{\Gamma})$ is $(\pi_{max}/(1+\pi_{max}), \bar{\pi}/(1+\bar{\pi}))$ where $\pi_{max} = \sup_i \pi_i^*(\mathbf{IQ})$ and $\bar{\pi} = (1/171)\sum_{i=1}^{171} \pi_i^*(\mathbf{IQ})$. It may concern some that $\pi_{max}/(1+\pi_{max})$ is a downwardly-biased estimator of Γ, but due to sampling variability and possible misspecification of the treatment and outcome models, the calibration is inherently approximate and meant only to act as a guide for the researcher conducting a sensitivity analysis of the Ashenfelter and Rouse study. It should also be noted that since higher IQ does not perfectly predict higher earnings, we find ourselves in a simultaneous sensitivity framework where we simultaneously bound the dependence between IQ and education and between IQ and earnings (see Gastwirth et al. (1998) for further details). This explains the slightly different definition of π_i^* used here than the one found in the paper. Simultaneous sensitivity analysis is closely related to amplified sensitivity analysis, which we discuss briefly in §3.7 of the paper (see Rosenbaum and Silber (2009) for more details). For our purposes, the simultaneous framework suffices to calibrate Γ and $\bar{\Gamma}$ in the Ashenfelter and Rouse study to the WLS study.

3.8.4. Details of Histogram in Right Panel of Figure 2

The figure in the right panel of Figure 2 in the paper is described as the *[h]istogram of the estimated increase in pairwise bias due to IQ disparities between siblings measured as an odds ratio.* To be specific, and using the notation introduced in Appendix 3.8.3, this is a histogram of

$$\frac{\pi_i^*(\mathbf{IQ})}{1 - \pi_i^*(\mathbf{IQ})} \Big/ \frac{\pi_i^*(\mathbf{0})}{1 - \pi_i^*(\mathbf{0})} \tag{3.22}$$

for $i = 1, \ldots, 171$ where $\pi_i^*(\mathbf{0}) = (1/2)$ is π_i^* computed for the sibling pair i had they had same IQ scores.

3.8.5. Additional Simulation Results

Table 9 shows the rejection probability of the true null hypothesis, $H_0 : \tau = 0$, under the *unbiased* setting with target Type I error control at $\alpha = 0.05$. Table 10 shows the power to reject the false null hypothesis, $H_0 : \tau = 0$, under the *unbiased* setting with true alternative hypothesis $H_1 : \tau = 0.25$ and target Type I error control at $\alpha = 0.05$.

			$\boldsymbol{\Gamma}$		
$\bar{\boldsymbol{\Gamma}}$	1	1.1	1.25	1.5	2
1	0.049	0.044	0.042	0.050	0.045
1.05		0.018	0.010	0.008	0.004
1.1		0.016	0.007	0.002	0.001
1.15			0.005	0.000	0.000
1.2			0.003	0.000	0.000
1.25			0.004	0.001	0.000
1.3				0.000	0.000
1.35				0.000	0.000
1.4				0.001	0.000
1.45				0.000	0.000
1.5				0.000	0.000
1.6					0.000
1.7					0.000
1.8					0.000
1.9					0.000
2					0.000

Table 9: Rejection probability of the true null hypothesis, $H_0 : \tau = 0$, under the *unbiased* setting with target Type I error control at $\alpha = 0.05$. The Monte Carlo standard error of these probability estimates is bounded above by $\sqrt{0.05 \times 0.95/5000} \approx 0.003$ if the true Type I error rate is 0.05.

	$\boldsymbol{\Gamma}$				
$\bar{\boldsymbol{\Gamma}}$	1	1.1	1.25	1.5	2
1	0.694	0.677	0.677	0.694	0.683
1.05		0.544	0.462	0.391	0.338
1.1		0.528	0.363	0.282	0.188
1.15			0.340	0.202	0.123
1.2			0.322	0.160	0.072
1.25			0.333	0.132	0.046
1.3				0.121	0.031
1.35				0.111	0.024
1.4				0.110	0.019
1.45				0.107	0.017
1.5				0.119	0.015
1.6					0.012
1.7					0.006
1.8					0.009
1.9					0.008
2					0.010

Table 10: Rejection probability of the false null hypothesis, $H_0 : \tau = 0$, under the *unbiased* setting with true alternative hypothesis $H_1 : \tau = 0.25$ and target Type I error control at $\alpha = 0.05$. The Monte Carlo standard error of these probability estimates is bounded above by $\sqrt{0.5 \times 0.5/5000} \approx 0.007$.

CHAPTER 4

Evaluating Missouri's Handgun Purchaser Law: A Bracketing Method for Addressing Concerns about History Interacting with Group

Abstract

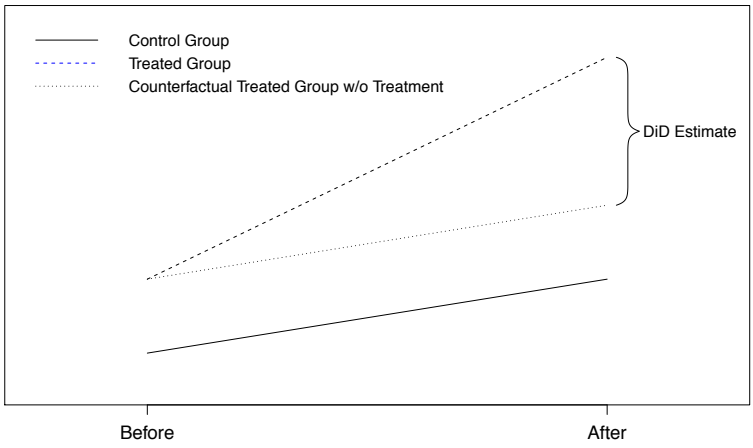
In the comparative interrupted time series design (also called the method of difference-in-differences), the change in outcome in a group exposed to treatment in the periods before and after the exposure is compared to the change in outcome in a control group not exposed to treatment in either period. The standard difference-in-difference estimator for a comparative interrupted time series design will be biased for estimating the causal effect of the treatment if there is an interaction between history in the after period and the groups; for example, there is a historical event besides the start of the treatment in the after period that benefits the treated group more than the control group. We present a bracketing method for bounding the effect of an interaction between history and the groups that arises from a time-invariant unmeasured confounder having a different effect in the after period than the before period. The method is applied to a study of the effect of the repeal of Missouri's permit-to-purchase handgun law on its firearm homicide rate. We estimate that the effect of the permit-to-purchase repeal on Missouri's firearm homicide rate is bracketed between 0.9 and 1.3 homicides per 100,000 people, corresponding to a percentage increase of 17% to 27% (95% confidence interval: [0.6,1.7] or [11%,35%]). A placebo study provides additional support for the hypothesis that the repeal has a causal effect of increasing the rate of state-wide firearm homicides.

4.1. Comparative Interrupted Time Series Design and Potential Biases

The interrupted time series design is an observational study design for estimating the causal effect of a treatment on a group when data is available before the group was treated. In the simplest interrupted time series design, the before and after treatment outcomes are compared. This before-after design does not account for confounding factors that co-occur with treatment such as historical events or maturation (Cook et al., 2002). To strengthen

the before-after design, it is common to add time series data from a control group that never received the treatment over the same period – the comparative interrupted time series design (Cook et al., 2002; Meyer, 1995; Bernal et al., 2017; Wing et al., 2018), also called the nonequivalent control group design or method of difference-in-differences. The latter name derives from the concept that the simplest comparative interrupted time series analysis is to take the difference between the difference of the after and before outcomes for the treated group and the difference of the after and before outcomes for the control group. This difference-in-differences estimate is an unbiased estimator of the causal effect of treatment if the treatment and control groups would have exhibited parallel trends in the counterfactual absence of treatment (Meyer, 1995); see Figure 5. The parallel trends

Figure 5: Stylized plot of data from a comparative interrupted time series design. The dotted line shows the assumption that the difference-in-difference (DiD) estimate makes about the treatment group's counterfactual mean in the absence of treatment.



assumption can be partially assessed if there is more than one time point in the before period by assessing whether the groups exhibit parallel trends in the before period (Meyer, 1995). However, even if the trends are parallel in the before period, there could be historical events in the after period that affect the two groups differently, i.e., history interacts with group (other reasons that parallel trends could be violated include differences in maturation, instrumentation or statistical regression between the groups) (Cook and Campbell, 1979; Reynolds and West, 1987). For example, the outcome measures poor health, country A

(treated group) enacts a policy reform, country B (control group) does not enact the reform, and a worldwide economic recession occurs after the reform that has a greater impact on people starting out in poorer health. If country B started out with poorer health, then parallel trends would be violated because country B's poor health would have increased more than country A in the after period in the counterfactual absence of the reform because of the worldwide economic recession. This violation of parallel trends would not happen if A and B started with the same level of poor health in the before period. However, it is often difficult to find a control group that has outcomes close to the treated group in the before period.

When there is no control group completely comparable to the treated group, Campbell (1969) proposed bracketing to distinguish treatment effects from plausible biases (Rosenbaum, 1987). Consider the study design of comparing treatment and control at one time point and suppose that there is concern about an unmeasured confounder U. Bracketing uses two control groups such that, in the first group U tends to be higher than in the treated group and in the second group, U tends to be lower. The effect of U on the treated group is bracketed by its effect on the two control groups. When there is bracketing, if the treated group has a notably higher outcome than both control groups, then this association between treatment and outcome cannot plausibly be explained away as being bias from U.

In this paper, we show how bracketing can be applied to the comparative interrupted time series to distinguish treatment effects from plausible biases due to history interacting with group. The basic idea is to consider one control group that has a lower expected outcome than the treated group in the before period and another control group that has a higher expected outcome than the treated group in the before period; we show under certain assumptions that the expectations of the two difference-in-difference estimators using the lower control group and higher control group respectively bracket the causal effect of the treatment. Bracketing for the comparative interrupted time series has been mentioned informally (Meyer, 1995) but the idea of choosing the bracketing control groups based on

expected before period outcomes was not mentioned. We present assumptions and results for our bracketing method in §4.2 and then apply the method to study the effect of the repeal of Missouri's permit-to-purchase handgun law on its firearm homicide rate in §4.3.

4.2. Methods: Bracketing

4.2.1. Notation and Model

Let Y denote outcome and D dose of exposure, $D = 1$ for treatment and $D = 0$ for control. Let $Y_{ip}^{(d)}$ denote the counterfactual outcome that would have been observed for unit i in period p, $p = 0$ for before period and $p = 1$ for after period, had the unit received exposure dose d, i.e., $Y_{ip}^{(1)}$ is the counterfactual outcome under treatment and $Y_{ip}^{(0)}$ is the counterfactual outcome under control. Let $\mathbf{U}_i$ be a vector of time invariant unmeasured confounders for unit i. Let G denote group where the groups are t = treated group, lc = lower control group (control group with expected outcomes lower than treated group in before period) and uc = upper control group (control group with expected outcomes higher than treated group in before period). Finally, let S be an indicator of whether or not a unit belonging to a particular group is in the study population in a given period. Specifically, $S_{ip} = 1$ or 0 when unit i is in the population or not in period p: $S_{i0} = S_{i1} = 1$ for a unit in the population both before and after treatment, $S_{i0} = 1, S_{i1} = 0$ for a unit in the population only before treatment (unit might have moved away or died in after period) and $S_{i0} = 0, S_{i1} = 1$ for a unit in the population only after treatment (unit might have moved into study area or been born in after period).

We consider the following model which generalizes the standard difference-in-difference model and changes-in-changes model. (Athey and Imbens, 2006) Let $\mathbf{U}_i$ be time-invariant unmeasured confounders and ϵ_{ip} be an error term that captures additional sources of variation for unit i in period p. Then our model can be expressed as

$$Y_{ip}^{(d)} = h(\mathbf{U}_i, p) + \beta d + \epsilon_{ip} \tag{4.1}$$

where the function $h(\mathbf{U}_i, p)$ is the unobserved expected outcome under control of subject i in period p. We drop the subscript i to refer to a randomly drawn unit from the population of all units in either period, where $Y_p^{(d)}, d = 0, 1$, and ϵ_p are undefined if $S_p = 0$. We make the following assumptions:

$$\textit{Increasingness of } h \textit{ in } \mathbf{U}\text{: } h(\mathbf{U}, p) \text{ bounded and increasing in } \mathbf{U} \text{ for } p = 0, 1. \tag{4.2}$$
$$((h(\mathbf{U}, p) \geq h(\mathbf{U}', p) \text{ whenever all coordinates of } \mathbf{U} \geq \text{all coordinates of } \mathbf{U}')$$

$$\textit{Time Invariance of } \mathbf{U} \textit{ Within Groups}\text{: } \mathbf{U} \text{ conditionally independent of } \{S_0, S_1\} \text{ given group } G. \tag{4.3}$$

$$\textit{Independence of } \epsilon \textit{ with Time and Group}\text{: Distributions of } \epsilon_p | S_p = 1, G = g \text{ for } p = 0, 1,\ g = lc, uc, tc \text{ all have mean zero and are the same.} \tag{4.4}$$

Assumptions (4.2) and (4.3) match assumptions in the changes-in-changes model. Assumption (4.2) requires that higher levels of unmeasured confounders correspond to higher levels of outcomes. Such increasingness is natural when the unmeasured confounder is an individual characteristic such as health or ability (Athey and Imbens, 2006) and Y is a measure of some positive outcome, for example, income. Negative confounders – where higher levels of the confounder correspond to lower levels of the outcome – are not precluded by Assumption (4.2) as the corresponding coordinates of $\mathbf{U}$ may simply be replaced by their negation. Assumption (4.3) says that the distribution of confounders in the population of units for a given group remains the same over time. Assumption (4.4) says that time-varying factors have the same distribution in each group and over time. It would be sufficient for subsequent developments to just assume the distributions of $\epsilon_p | S_p = 1, G = g$ for $p = 0, 1,\ g = lc, uc, tc$ all have mean zero rather than the stronger assumption of identical distributions. We can further relax this assumption by assuming zero mean only for components of ϵ_p that are true confounders, that is, factors whose distributions depend on the interaction of time and

group. Assumption (4.4) is weaker than the changes-in-changes model assumption that ϵ_{ip} is always zero which rules out classical measurement error in the outcome when h is non-linear (Athey and Imbens, 2006). Our model contains the standard difference-in-difference model, which can be represented in our model by $h(\mathbf{U}, p) = k(\mathbf{U}) + \tau p$ for some bounded and increasing function k, where $k(\mathbf{U})$ can be viewed as a group fixed effect.

We make two further assumptions about the distribution of $\mathbf{U}$ in groups and how its effect over time changes among the groups. First, we assume the distribution of $\mathbf{U}$ within groups can be stochastically ordered so that $\mathbf{U}$ is lowest in the lower control group, intermediate in the treated group and highest in the upper control group:

$$\mathbf{U}|G = lc \preceq \mathbf{U}|G = t \preceq \mathbf{U}|G = uc \tag{4.5}$$

where two random vectors $\mathbf{A}, \mathbf{B}$ are stochastically ordered, $\mathbf{A} \preceq \mathbf{B}$, if $E[f(\mathbf{A})] \leq E[f(\mathbf{B})]$ for all bounded increasing functions f (Shaked and Shanthikumar, 1994). For example, if $\mathbf{U}$ is normally distributed with common variance and group means μ_{lc}, μ_t, and μ_{uc}, then $\mu_{lc} \leq \mu_t \leq \mu_{uc}$ would imply (4.5). Second, we assume that higher values of $\mathbf{U}$ either have a bigger effect over time over the whole range of $\mathbf{U}$ or a smaller effect over the whole range:

$$\begin{aligned} \text{Either} \quad & \text{(i) } h(\mathbf{U}, 1) - h(\mathbf{U}, 0) \geq h(\mathbf{U}', 1) - h(\mathbf{U}', 0) \text{ for all } \mathbf{U} \geq \mathbf{U}', \ \mathbf{U}, \mathbf{U}' \in \mathcal{U} \quad \text{or} \\ & \text{(ii) } h(\mathbf{U}, 1) - h(\mathbf{U}, 0) \leq h(\mathbf{U}', 1) - h(\mathbf{U}', 0) \text{ for all } \mathbf{U} \geq \mathbf{U}', \ \mathbf{U}, \mathbf{U}' \in \mathcal{U} \end{aligned} \tag{4.6}$$

An example of this pattern of $\mathbf{U}$ confounding could occur in a study of the effect of a regional policy on average income where the policy change occurred contemporaneously with an easing of trade restrictions. A potential unmeasured confounder for such a study would be $\mathbf{U}$ = share of skilled workers in a region, as a higher share of skilled workers is associated with higher average income. There is considerable evidence that trade liberalization leads to an increase in the skill premium – the relative wage of skilled to unskilled workers – at both the regional and country level (Dix-Carneiro and Kovak, 2017; Burstein and Vogel,

2017). Thus, we might expect (i) in (4.6) to hold if there was an easing of trade restrictions in the after period.

We assume units are randomly sampled from each group in each time period. The data could be obtained from repeated cross sections or a longitudinal study. Inferences under different sampling assumptions are discussed in Appendix 4.5.1.

4.2.2. Bracketing Result

The standard moment difference-in-difference estimator using control condition c can be written as $\hat{\beta}_{dd.c} = (\overline{Y}_{1|G=t} - \overline{Y}_{0|G=t}) - (\overline{Y}_{1|G=c} - \overline{Y}_{0|G=c})$ where $\overline{Y}_{p|G=g}$ indicates the sample average of units observed in group g and time period p, $Y_p|G = g, S_p = 1$. This estimate is equivalent to the coefficient on the treatment indicator in a fixed effects regression with full time and group indicator variables. When using data already aggregated at some level, for example by state-year, a fixed effects regression using weights proportional to population will return this estimate. In the following, we show that the expectation of the two standard difference-in-difference estimators computed with the upper and lower controls can be used to bound the treatment effect.

The expected value of the standard difference-in-difference estimator comparing the treated group to the lower control group, $\hat{\beta}_{dd.lc}$, is

$$\begin{aligned} E[\hat{\beta}_{dd.lc}] &= \{E[Y_1|G=t, S_1=1] - E[Y_0|G=t, S_0=1]\} \\ &\quad -\{E[Y_1|G=lc, S_1=1] - E[Y_0|G=lc, S_0=1]\} \\ &= \{\beta + E[h(\mathbf{U},1)|G=t, S_1=1] - E[h(\mathbf{U},0)|G=t, S_0=1]\} \\ &\quad -\{E[h(\mathbf{U},1)|G=lc, S_1=1] - E[h(\mathbf{U},0)|G=lc, S_0=1]\}, \end{aligned}$$

where Y_1, Y_0 denote observed outcomes in after period ($p = 1$) and before period ($p = 0$) respectively. Under the time invariance of $\mathbf{U}$ within groups assumption (4.3), we have

$$E[\hat{\beta}_{dd.lc}] = \beta + \{E[h(\mathbf{U},1) - h(\mathbf{U},0)|G=t]\} - \{E[h(\mathbf{U},1) - h(\mathbf{U},0)|G=lc]\}; \qquad (4.7)$$

similarly, the expected value of the difference-in-difference estimator comparing the treated group to the upper control group, $\hat{\beta}_{dd.uc}$, is

$$E[\hat{\beta}_{dd.uc}] = \beta + \{E[h(\mathbf{U},1) - h(\mathbf{U},0)|G = t]\} - \{E[h(\mathbf{U},1) - h(\mathbf{U},0)|G = uc]\}. \quad (4.8)$$

The difference-in-difference estimators $\hat{\beta}_{dd.lc}$ and $\hat{\beta}_{dd.uc}$ are unbiased if $h(\mathbf{U},1) - h(\mathbf{U},0)$ is constant for all $\mathbf{U}$, i.e., the time effect between periods is the same for all levels of $\mathbf{U}$, or equivalently, the effect of the unmeasured confounders is the same in both time periods. If the effect of the unmeasured confounders changes between periods, then because of assumptions (4.5) and (4.6), we conclude from (4.7) and (4.8) that

$$\min\{E[\hat{\beta}_{dd.lc}], E[\hat{\beta}_{dd.uc}]\} \leq \beta \leq \max\{E[\hat{\beta}_{dd.lc}], E[\hat{\beta}_{dd.uc}]\}, \quad (4.9)$$

i.e., the expected values of the difference-in-difference estimators using the upper control group and lower control group bracket the causal effect (proof in Appendix 4.5.2). The tightness of the bracketing bounds in (4.9) and, to some extent, the width of the corresponding confidence interval developed in following section depend on the magnitude of the group-by-time interaction. For example, if urban poverty concentration varied notably between groups and its effect on firearm homicides were modulated by the Great Recession, one would expect looser bracketing bounds.

4.2.3. Inference

We would like to make inferences for the causal effect β under the assumption (4.6) that $h(\mathbf{U},1) - h(\mathbf{U},0)$ is either an increasing or decreasing function of $\mathbf{U}$ (we do not want to specify which a priori). Let $\theta_{lc.t} = E[\hat{\beta}_{dd.lc}]$ and $\theta_{uc.t} = E[\hat{\beta}_{dd.uc}]$, i.e., the expected values of the difference-in-difference estimators using the lower control group and upper control group, respectively. From the bracketing results (4.9), we have

$$\min(\theta_{lc.t}, \theta_{uc.t}) \leq \beta \leq \max(\theta_{lc.t}, \theta_{uc.t}).$$

and the following interval, where CI means two-sided confidence interval,

$$[\min(\text{lower bound of } 1-\alpha \text{ CI for } \theta_{lc.t}, \text{ lower bound of } 1-\alpha \text{ CI for } \theta_{uc.t}),$$
$$\max(\text{upper bound of } 1-\alpha \text{ CI for } \theta_{lc.t}, \text{ upper bound of } 1-\alpha \text{ CI for } \theta_{uc.t})], \qquad (4.10)$$

has probability $\geq 1-\alpha$ of containing both $\min(\theta_{lc.t}, \theta_{uc.t})$ and $\max(\theta_{lc.t}, \theta_{uc.t})$, and thus β, where it assumed that the two-sided CIs are constructed by taking the intersection of two one-sided $1-(\alpha/2)$ confidence intervals (proof in Appendix 4.5.3).

4.2.4. Constructing the Lower and Upper Control Groups

The results in §4.2.2-4.2.3 assume the lower and upper control groups have been constructed before looking at the data. If the lower control group was constructed by looking at the before period data by choosing units with lower outcomes than the treated in the before period, then the sample average of $Y_0|G = lc, S_0 = 1$ may tend to be lower than $E(Y_0|G = lc, S_0 = 1)$. Consequently, the difference-in-difference estimate using the lower control group may be downward biased even if the parallel trends assumption holds because of regression to the mean (Cook et al., 2002); similarly, the difference-in-difference estimated using the upper control group may be upward biased. This may invalidate the bracketing result (4.9). To avoid bias arising from regression to the mean, we propose first selecting a "pre-study" time period prior to the before period. Then, the lower control group can be constructed from units with lower outcomes than the treated in this pre-study period and the upper control group from units with higher outcomes. It should then be tested whether the constructed lower control group has smaller expected outcomes than the constructed upper control group in the before period; see §4.3 for example.

4.2.5. Role of Examining the Groups' Relative Trends in the Before Period

In the standard difference-in-difference analysis that assumes parallel trends, when the before period contains multiple time points, it is good practice to test for parallel trends in the before period (Meyer, 1995; Volpp et al., 2007). In our bracketing approach, we do not need

the parallel trend assumption to hold, but examining the relative trends of the groups in the before period is still useful for assessing model plausibility and assumptions. Our model (4.1)-(4.4) along with assumptions (4.5)-(4.6) implies that if we had counterfactual data on the treatment group in the after period in the absence of treatment, then, without sampling variance, we would see either: (i) the differences between the upper control and counterfactual treated groups and the difference between the counterfactual treated and lower control groups in the after period would be at least as large as their respective differences in the before period or (ii) the difference between the upper control and counterfactual treated groups and the difference between the counterfactual treated and lower control groups in the after period would be no larger and possibly smaller than their respective differences in the before period. The following two patterns would violate the model/assumptions: (iii) the difference between the upper control and counterfactual treated groups is larger after than before and the difference between the counterfactual treated and lower control groups is smaller after than before or (iv) the difference between the upper control and counterfactual treated groups is smaller after than before and the difference between the counterfactual treated and lower control groups is larger after than before. Although we do not have the counterfactual treatment group's data in the absence of treatment in the after period, we have the treatment group's data in the absence of treatment in the before period. We can split the before period into two (or more) periods and test whether the pattern in the before period is consistent with the model. Visual inspection of the relative trends of the counterfactual treated group and the upper and lower control groups during the before period can provide additional evidence for or against the model assumptions.

4.2.6. Time-Varying Confounders

Our bracketing method addresses an interaction between history and groups that arises because the time-invariant unmeasured confounders that differ between the groups in the before period ($\mathbf{U}$) become more (or less) important in the after period (assumption (4.6)). When there are time-varying confounders, the bracketing method still works under certain

assumptions. Time-varying confounders can be represented in model (4.1) by letting $\mathbf{U}$ contain all variables that differ in distribution between the groups in the before period, ϵ_{i0} be the effect of factors that do not differ in distribution between the groups in the before period and ϵ_{i1} be the effect of the same factors in ϵ_{i0} in the after period as well as factors not contained in $\mathbf{U}$ that differ in distribution between the groups in the after period (details on time-varying model in Appendix 4.5.4. If this last set of factors is present, then (4.4) may not hold. However, the bracketing result (4.9) still holds as long as (i) in (4.6) holds,

$$E[\epsilon_{i1}|G = uc] \geq E[\epsilon_{i1}|G = t] \geq E[\epsilon_{i1}|G = lc], \tag{4.11}$$

or when (ii) in (4.6) holds,

$$E[\epsilon_{i1}|G = uc] \leq E[\epsilon_{i1}|G = t] \leq E[\epsilon_{i1}|G = lc]; \tag{4.12}$$

Appendix 4.5.4 contains a proof and sufficient conditions for (4.11) or (4.12) to hold. One of these sufficient conditions (condition (c) in Appendix 4.5.4 is analogous to (i) in (4.6) in that effects on the outcome, be they time effects or those due to contemporaneous shocks to confounders, are amplified at larger values of $\mathbf{U}$.

One type of time-varying confounder is a variable that largely stays the same between time periods but may change modestly. For example, in our study of Missouri's repeal of their permit-to-purchase law in §4.3, urban concentration of poverty might be a confounder and $\mathbf{U}$ contain urban concentration of poverty in the before period. Urban concentration of poverty may stay mostly the same over time but change modestly, where the changes are reflected in ϵ_1. If the effect of urban concentration of poverty on firearm homicides increased in the after period, then the bracketing result would still hold (with respect to the confounding from urban concentration of poverty) as long as the impact of changes in urban concentration of poverty on firearm homicides were at least as great in the upper control group as Missouri and at least as great in Missouri as the lower control group.

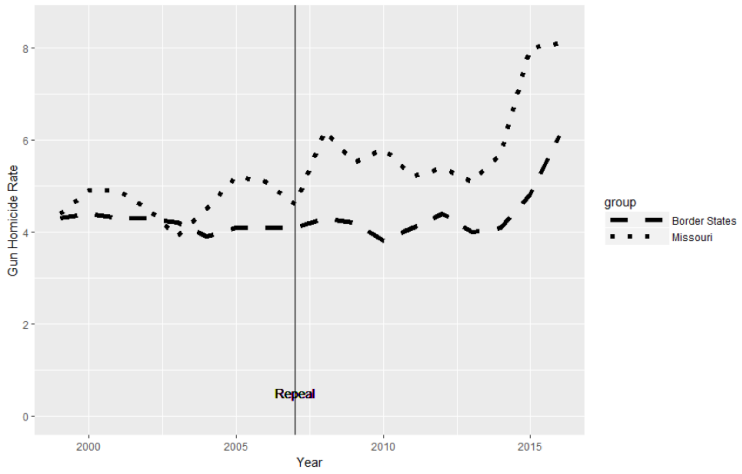
4.3. Application: Effect of the Repeal of Missouri's Handgun Purchaser Licensing Law on Firearm Homicides

American federal gun law requires background checks and record keeping for gun sales by federally licensed firearm dealers but exempts these regulations for private sales. However, some states have laws requiring all purchasers of handguns from licensed dealers *and* private sellers to acquire a permit-to-purchase license that verifies the purchaser has passed a background check. Missouri passed a permit-to-purchase law in 1921, requiring handgun purchasers to obtain a license from the local sheriff's office that facilitated the background check, but repealed the law on August 28, 2007. Webster et al. (2014) examined the effect of Missouri's repeal on firearm homicide rates (the rate of homicides committed using a firearm). One of their analyses used a comparative interrupted time series design, comparing Missouri to the eight states bordering Missouri using a before-period of 1999-2007 and after-period of 2008-2010 (the only available post-repeal data at the time of their analysis), finding evidence that the repeal of Missouri's permit-to-purchase law increased firearm homicide rates (see their Table 1). None of the border states introduced new or made changes to existing permit-to-purchase laws during the study period. Using a fixed effect regression and adjusting for several background crime and economic covariates, they estimated that the Missouri permit-to-purchase repeal was associated with an increase in the firearm homicide rate by 1.1 per 100,000 persons (95% confidence interval [CI]: 0.8,1.4) , a 22% (95% CI: 16 %, 29%) increase. Non-gun related homicides remained virtually unchanged. In what follows, we re-examine the effect of Missouri's repeal using bracketing and the now available after-period data from 2008-2016 to address possible biases arising from unobserved state-by-time interactions.

Figure 6 shows the age-adjusted firearm homicide rates in Missouri and the border states over the study period using data from the Centers for Disease Control and Prevention (CDC) Wide-ranging Online Data for Epidemiologic Research (WONDER) system (http://wonder.cdc.gov, 2018). The standard difference-in-difference estimate using all neighboring control states,

shown in the top row of Table 12, is that Missouri's permit-to-purchase repeal increased firearm homicides by 1.2 per 100,000 persons (95% CI: 1.0,1.4), corresponding to a 24% increase (95% CI: 18%,31%). In the before-period, Missouri had generally higher firearm

Figure 6: Age-adjusted firearm homicide rates in Missouri and states bordering Missouri (population-weighted averages), 1999-2016.



homicide rates than the control border states, suggesting a lack of comparability between the groups. One concern is that the start of the after period coincided with the beginning of the Great Recession. The economic downturn was followed by a decline in homicide rates. Possible reasons for the effect of the downturn on homicide rates and violence generally include changing alcohol affordability, disposable income, unemployment, and income inequality (Matthews et al., 2006; Wolf et al., 2014; Shepherd and Page, 2015). The effects of the economic downturn on firearm homicides might interact with the starting level of firearm homicides in a state. To address this concern, we constructed upper and lower control groups that bracket Missouri's firearm homicide rate in the before period. To avoid regression to the mean in §4.2.4, we use data from 1994-1998, the five years prior to our before period, to choose the upper and lower control groups; see Table 11 for data. The lower control group is Iowa, Kansas, Kentucky, Nebraska, and Oklahoma and the upper control group is Arkansas, Illinois, and Tennessee. The population-weighted firearm homicide rate in the before period of 1999-2007 is 5.2 in the upper control states, 4.7 in Missouri, and

2.7 in the lower control states (95% CI for difference between upper control and Missouri: 0.2,0.8; 95% CI for difference between Missouri and lower controls: 1.8,2.2).

Table 11: Age-adjusted firearm homicide rates per 100,000 persons from periods 1994-1998 (pre-study period used to construct lower and upper control groups), 1999-2007 (before repeal period where repeal refers to repeal of Missouri's permit-to-purchase handgun licensing law) and 2008-2016 (after repeal period).

	1994-1998	1999-2007	2008-2016
Missouri	6.1	4.7	6.1
Arkansas	7.3	5.1	5.5
Illinois	7.1	5.1	5.2
Iowa	1.2	0.9	1.2
Kansas	4.2	3.0	3.0
Kentucky	4.1	3.3	3.7
Nebraska	2.2	1.8	2.4
Oklahoma	4.8	3.8	4.8
Tennessee	6.9	5.5	5.4
Population-weighted All Controls	5.6	4.2	4.4
Population-weighted Upper Controls	7.1	5.2	5.3
Population-weighted Lower Controls	3.5	2.7	3.2

Figure 7 shows firearm homicides rates (age-adjusted and population-weighted) in the bracketed control groups compared to Missouri. The bottom two rows of Table 12 show the difference-in-difference estimates using the lower and upper control groups and 95% CIs. Both the lower and upper control groups provide evidence that Missouri's repeal of its permit-to-purchase handgun law increased firearm homicides, bracketing the effect of the repeal between 0.9 and 1.3 homicides per 100,000 people, corresponding to a percentage increase of 17% to 27%. The interval (4.10) that has a $\geq 95\%$ chance of containing the effect of the repeal on the firearm homicide rate is $[0.6, 1.7]$, corresponding to an 11% to 35% increase in firearm homicides, providing evidence that the repeal increased firearm homicides.

4.3.1. Assessing Model Assumptions: Time-Varying Confounders and Relative Trends

A type of time-varying confounder that is relevant to the Missouri permit-to-purchase study

Figure 7: Age-adjusted gun homicide rates per 100,000 persons in Missouri, lower control states bordering Missouri (population-weighted averages) and upper control states bordering Missouri, 1999-2016.

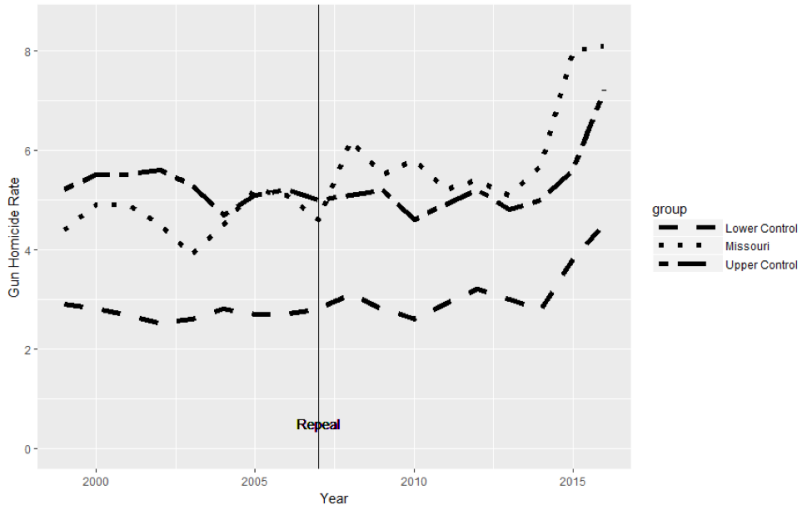


Table 12: Difference-in-difference estimates of effect of repeal of Missouri's permit-to-purchase handgun licensing requirement on firearm homicide rates per 100,000 persons. CI indicates confidence interval.

Control Group	Estimate	95% CI	% Change Estimate	95% CI
All Controls	1.2	[0.9, 1.5]	24%	[18% ,31%]
Upper Controls	1.3	[0.9, 1.7]	27%	[19% ,35%]
Lower Controls	0.9	[0.6, 1.2]	17%	[11% ,23%]

is a factor that only arises in the after period. The Ferguson unrest in 2014 might have led to less effective policing (spikes in violence typically follow social unrest) in Missouri compared to other states. Such a time-varying confounder would be unlikely to satisfy (4.11) or (4.12) because it arises only in the treated group (Missouri) in the after period. However, this confounder alone does not change our finding that the repeal increased firearm homicides. If we limit the study to 2008-2013, Missouri still has larger increases in firearm homicide rates than both the upper and lower control groups; see Appendix 4.5.6.

To assess the plausibility of our model (4.1)-(4.4) and assumptions (4.5)-(4.6), we apply the relative trends test described in §4.2.5. Applying the test to our study of the repeal of Missouri's permit-to-purchase law, we do not find evidence that our model assumptions

are violated. Visual inspection of the relative trends of counterfactual Missouri and the upper and lower controls in the before period further supports the plausibility of our model assumptions; see Figure 9 in Appendix 4.5.5.

4.3.2. Standard Error Estimates: A Poisson Model for Death Counts

The standard errors used for inference in the previous section come directly from the CDC WONDER system. Vital statistics that derive from complete counts of deaths (by cause) are not subject to sampling error. Nonetheless, a stochastic model of vital statistics may be justified by the presence of biological, environmental, sociological, and other natural sources of variability (Brillinger, 1986). For inferential purposes, a census may be viewed as a realization from such a stochastic process under similar conditions to those observed (Keyfitz, 1966). In particular, the observed firearm homicide death rate in any state-year may be viewed as one of a large series of possible Poisson distributed outcomes under similar conditions (US Department of Health and Human Services, 2004). The standard errors reported by the CDC are computed under this Poisson model.

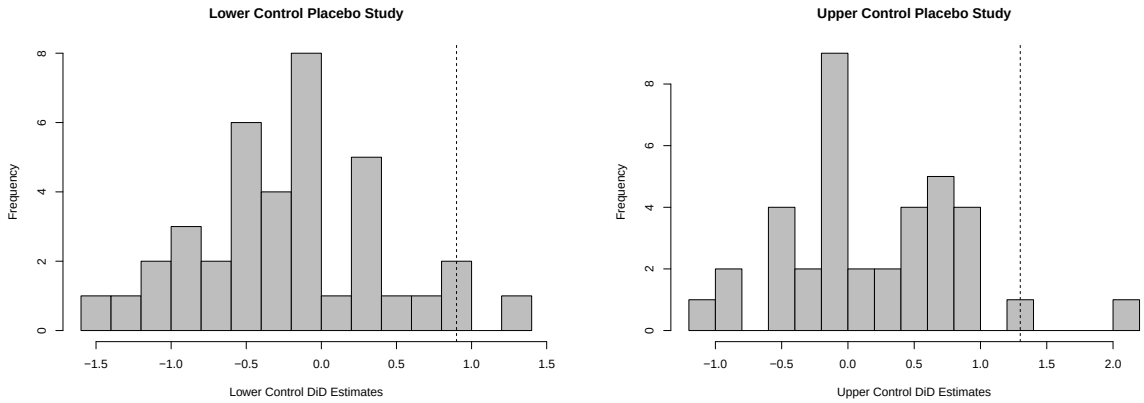
4.3.3. A Placebo Study: Assessing Alternative Sources of Uncertainty

There may be other sources of uncertainty unaccounted for by the natural variability of a Poisson model for yearly state-level firearm homicides. Several recent papers suggest that such sources of uncertainty, if ignored, may yield substantially different inferential conclusions. Serially correlated data (Bertrand et al., 2004), yearly state-level shocks (Donald and Lang, 2007), and small numbers of policy changes (Conley and Taber, 2011) can cause the standard errors returned by a fixed effects regression to be downwardly biased. We conduct a placebo study (Abadie et al., 2010; Bertrand et al., 2004) to address inferential challenges that arise from the presence of possibly dependent, yearly state-level shocks to the conditions that generate these Poisson realizations.

Akin to permutation inference, a placebo study in the context of the Missouri permit-to-purchase repeal analysis applies the bracketing method to every state to create a placebo

intervention effect distribution. Specifically, for each state where there was no permit-to-purchase repeal we construct lower and upper control groups of neighboring states, when available, in exactly the same way we did so for Missouri. We then compute the difference-in-difference estimates using both control groups for a placebo "repeal" on August 28, 2007. This results in two exact distributions for the placebo intervention effect estimate, one estimated using lower controls and the other using upper controls. If the permit-to-purchase repeal effect in Missouri is not spurious, we would expect to see few placebo effects greater than the ones reported in our study using either control condition. The

Figure 8: Histograms of placebo "repeal" effects using different control states. (Left Panel): Histogram of placebo difference-in-difference estimates using lower control states ($n = 38$ states with lower control neighbors – includes Missouri). Two states (Oklahoma and Delaware) had a larger estimate than Missouri (dashed line).(Right Panel): Histogram of placebo difference-in-difference estimates using upper control states ($n = 37$ states with upper control neighbors – includes Missouri). One state (Delaware) had a larger estimate than Missouri (dashed line).



histograms of the placebo effects in Figure 8 suggest that the Missouri bracketing study is relatively robust to these alternative sources of variability. Of the 38 states that had lower control neighbors, only two (Oklahoma and Delaware) had placebo effect estimates using lower controls that were larger than Missouri (dashed line, left panel). Of the 37 states that had upper control neighbors, only one (Delaware) had a placebo effect estimate using upper controls that was larger than Missouri (dashed line, right panel). Alaska, Hawaii, the District of Columbia and three states with missing data in either the pre-study, before or

after period were excluded from the analysis.

4.4. Conclusion and Discussion

We developed a bracketing method for comparative interrupted time series to account for concerns that history may interact with groups. In a study of the repeal of Missouri's permit-to-purchase handgun law, the method addressed a concern that on average, control states started out with lower firearm homicide rates than Missouri before the repeal. Comparing both to states that started with higher firearm homicide rates than Missouri and states that started with lower rates, the repeal was associated with a significant increase in firearm homicides, thus strengthening the evidence that the repeal had a causal effect of increasing firearm homicides.

A limitation of our estimated impact of the repeal of Missouri's permit-to-purchase law is that a Stand Your Ground law was simultaneously adopted in Missouri. However, in the original study by Webster et al. (2014), the inclusion of a Stand Your Ground indicator in the regression did not dramatically change the estimated effect. Additionally, a recent comparative interrupted time series study examining firearm homicide rates in large urban counties found that permit-to-purchase laws were associated with significant reductions in firearm homicides after controlling for the effects of Stand Your Ground lawsCrifasi et al. (2018). Further evidence that the contemporaneous Stand Your Ground law does not change the qualitative conclusion of our study can be found in the placebo study. There were 16 additional states that adopted Stand Your Ground laws within a few years of Missouri's permit-to-purchase repealCrifasi et al. (2018). Only one state (Oklahoma) of the 16 had a difference-in-difference placebo effect estimate using lower controls that was larger than Missouri and none of the states had placebo effect estimates using upper controls that were larger than Missouri.

Although only one of many potential patterns of bias, the history-by-group interaction bias addressed in this paper has been mentioned in the literature since at least the middle of the

20th century. A version of it is referred to *selection-maturation interaction* in a taxonomy of possible threats to the validity of experimental and quasi-experimental designs presented in Campbell and Stanley (Campbell and Stanley, 1963). Fundamentally, bracketing relies on constructing control groups across which this potential source of confounding is systematically varied (Hasegawa and Small, 2017). Other methods for constructing adequate control groups in the presence of history-by-group interactions, such as the synthetic control method (Abadie et al., 2010), have also found success in comparative case studies of the effect of permit-to-purchase laws on firearm homicide rates (Rudolph et al., 2015). While we do not argue that bracketing is uniformly superior to the synthetic control method, the practitioner may find that each has strengths that lend themselves to different settings. When the researcher believes that unmeasured history-by-group confounding, $h(\mathbf{U}, p)$, can be expressed as a linear factor model with time-varying slopes and group-specific loadings, the synthetic control method provides an asymptotically unbiased point estimate of the causal effect of treatment while bracketing can only provide bounds on the treatment effect. However, when the practitioner suspects that only the weaker assumptions of the model outlined in §4.2.1 hold, the bracketing bounds will remain unbiased, in that they contain the true effect in expectation, while the point estimate using synthetic controls need not be unbiased; see Appendix 4.5.7 for further discussion. A detailed example of such a case can be found in the Appendix 4.5.8.

4.5. Appendix

4.5.1. Inferences Under Different Sampling Assumptions

The standard difference-in-difference estimator using a control group c, $\hat{\beta}_{dd.c}$, is

$$\begin{aligned}\hat{\beta}_{dd.c} &= \{\hat{E}[Y_1|G=t, S_1=1] - \hat{E}[Y_0|G=t, S_0=1]\} \\ &\quad -\{\hat{E}[Y_1|G=c, S_1=1] - \hat{E}[Y_0|G=c, S_0=1]\}.\end{aligned}$$

When the samples of (i) $Y_1|G=t,S_1=1$, (ii) $Y_0|G=t,S_0=1$, (iii) $Y_1|G=c,S_1=1$ and (iv) $Y_0|G=c,S_0=1$ are independent, then the standard error of $\hat{\beta}_{dd.c}$ is

$$SE(\hat{\beta}_{dd.c}) =$$
$$\{SE(\hat{E}[Y_1|G=t,S_1=1])^2 + SE(\hat{E}[Y_0|G=t,S_0=1])^2$$
$$+SE(\hat{E}[Y_1|G=c,S_1=1])^2 + SE(\hat{E}[Y_0|G=c,S_0=1])^2\}^{1/2}. \quad (4.13)$$

We use (4.13) to make inferences for our study of the effect of the repeal of Missouri's permit-to-purchase law, where the $\hat{E}$ and corresponding SEs are obtained from the CDC's WONDER system.

Let κ_{tt} be the % change in the treated group's mean outcome in the after period compared to its mean counterfactual outcomes in the after period in the absence of treatment,

$$\kappa_{tt} = 100 \times \frac{E[Y_1^{(1)}|G=t,S_1=1] - E[Y_1^{(0)}|G=t,S_1=1]}{E[Y_1^{(0)}|G=t,S_1=1]}.$$

An estimate of κ_{tt} using control group c and assuming the parallel trends of standard-in-differences is

$$\hat{\kappa}_{tt.c} = 100 \times \frac{\hat{\beta}_{dd.c}}{\hat{E}(Y_0|G=t,S_0=1) + \{\hat{E}(Y_1|G=c,S_1=1) - \hat{E}(Y_0|G=c,S_0=1)\}}.$$

We approximate the standard error of $\hat{\kappa}_{tt.c}$ using the Delta method.

The model (4.1) can be extended to allow for observed covariates, clustering and multiple time points using a regression framework (Imbens and Wooldridge, 2009). The difference-in-difference estimator may be computed by regressing the observed outcome Y on a time period dummy, a group dummy and a treatment variable. Observed covariates $\mathbf{X}_{ip}$ that could vary by time can be incorporated into the model and then the difference-in-difference regression estimator can be computed by regressing Y on the observed covariates, a time period dummy, a group dummy and a treatment variable. The model assumptions then

need to hold only conditionally on the observed covariates. The comparative interrupted time series can be applied to settings with more than two time periods. A full set of time period dummies can be added to model (4.1). The effect of the treatment over time can be allowed to vary by interacting the treatment dummy with time.

Within each group, there may be clusters of units, e.g., different countries that had the same policy reform. For such settings, we can extend model (4.1) to the following (Donald and Lang, 2007) where the index cip denotes the ith unit in cluster c at time period p:

$$Y_{cip}^{(d)} = h(\mathbf{U}_{cip}, p) + \beta d + \eta_{cp} + \epsilon_{cip}, \tag{4.14}$$

where η_{cp} represents an effect shared by members of cluster c in period p, e.g., an economic shock that is specific to a country c in period p. Under an assumption that the η_{cp} are independent and identically distributed (i.i.d.) normal random variables, Donald and Lang (2007) showed that if we compute the mean in each cluster at each time period, and regress these cluster/period means on fixed effects for each cluster, a time period dummy and a treatment variable, then the t statistic for the treatment variable ($\frac{\hat{\beta}-\beta}{SE(\hat{\beta})}$) has a t distribution with the number of clusters minus two degrees of freedom. Using this approach, we do not need to have individual data but only summary data for each cluster. Other approaches to inference that allow for the η_{cp} to be non-i.i.d. such as autocorrelated within group, have been developed. (Bertrand et al., 2004; Hansen, 2007).

Note that the presence of at least two clusters in at least one group enables us to make inferences that allow for shared effects η_{cp}. When there is only one cluster in each group, e.g., we are comparing just two countries, one in which a policy reform was implemented and one in which it was not, then there are zero degrees of freedom to estimate the variance of the η_{cp} so inferences cannot be drawn that allow for η_{cp} to be nonzero using data from entirely within the sample. For such settings, it may be possible to get information from outside the sample to get a plausible estimate of the variance of the η_{cp} (Blitstein et al., 2005; Donald and Lang, 2007).

4.5.2. Proof of (4.9) *in §4.2.2*

Suppose $h(\mathbf{U},1) - h(\mathbf{U},0)$ is a bounded increasing function of $\mathbf{U}$. Then from (4.5) and the property that bounded increasing functions of stochastically ordered random variables preserve order, it follows that

$$E[\hat{\beta}_{dd.uc}] \leq \beta \leq E[\hat{\beta}_{dd.lc}]. \tag{4.15}$$

Similarly, if $h(\mathbf{U},1) - h(\mathbf{U},0)$ is a bounded decreasing function of $\mathbf{U}$,

$$E[\hat{\beta}_{dd.lc}] \leq \beta \leq E[\hat{\beta}_{dd.uc}]. \tag{4.16}$$

(4.9) follows from (4.15) and (4.16).

4.5.3. Proof for Result in §4.2.3

Here we prove that (4.10) has probability $\geq 1-\alpha$ of containing both $\min(\theta_{lc.t}, \theta_{uc.t})$ and $\max(\theta_{lc.t}, \theta_{uc.t})$ under the assumption that the two sided CIs are constructed in the usual way by taking the union of two one-sided $1-(\alpha/2)$ confidence intervals. The result is basically derived by inverting multiparameter hypothesis tests about the minimum or maximum of two parameters (Lehmann, 1952; Berger, 1982). Let $q = \min(\theta_{lc.t}, \theta_{uc.t})$ and $r = \max(\theta_{lc.t}, \theta_{uc.t})$. The probability that (4.10) does not contain both $\min(\theta_{lc.t}, \theta_{uc.t})$ and $\max(\theta_{lc.t}, \theta_{uc.t})$ is bounded by the probability that q is less than the lower endpoint of the interval plus the probability that r is greater than the upper endpoint of the interval. The probability that q is less than the lower endpoint of the interval is the probability that both one-sided tests $H_0^l : \theta_{lc.t} \leq q$ vs. $H_1^l : \theta_{lc.t} > q$ and $H_0^u : \theta_{uc.t} \leq q$ vs. $H_1^l : \theta_{uc.t} > q$ give p-values $\leq \alpha/2$, which has probability at most $\alpha/2$ since each individual event has probability at most $\alpha/2$. Similarly, the probability that r is greater than the upper endpoint of the interval is the probability that both one-sided tests $H_0^{l'} : \theta_{lc.t} \geq r$ vs. $H_1^{l'} : \theta_{lc.t} < r$ and $H_0^{u'} : \theta_{uc.t} \geq r$ vs. $H_1^{l'} : \theta_{uc.t} < r$ give p-values $\leq \alpha/2$, which has probability at most $\alpha/2$

since each individual event has probability at most $\alpha/2$. Thus, the probability that (4.10) does not contain both $\min(\theta_{lc.t}, \theta_{uc.t})$ and $\max(\theta_{lc.t}, \theta_{uc.t})$ is bounded by α.

4.5.4. Modeling Time-varying Confounders

We model a setting with time-varying confounders as follows. We maintain the assumptions in §4.2.1 except for (4.4). We let $\mathbf{U}$ contain all variables that affect the outcome that differ in distribution between the groups (treated, upper control, lower control) in the before period and let ϵ_0 summarize the effect of factors in the before period that do not differ in distribution between the groups. We can model the average effect of the factors in ϵ_0 as an intercept in the $h(\mathbf{U}, 0)$ function so that $E(\epsilon_0|S_0 = 1, G = g) = 0$ holds for all groups $g = lc, uc, tc$. The effect of factors that do not differ in distribution between the groups in the after period as well as the effect of time-varying confounders in the after period are summarized in ϵ_1. Some of these time-varying confounders may be variables in $\mathbf{U}$ that have changed their level over time. Let $\mathbf{U}_0 \equiv \mathbf{U}$ be the value of the variables in $\mathbf{U}$ in the before period and $\mathbf{U}_1$ be their value in the after period, where $\mathbf{U}_0 = \mathbf{U}_1$ for a unit only in the population in the after period (with $\mathbf{U}$ defined this way, the validity of (4.3) needs to be considered carefully). Then, assuming that the average effect of the factors in ϵ_1 that do not differ between the groups in the after period is modeled as an intercept in $h(\mathbf{U}, 1)$, we have

$$E(\epsilon_1|G = g, S_1 = 1) = E[h(\mathbf{U}_1, 1) - h(\mathbf{U}_0, 1)|G = g, S_1 = 1].$$

Then for (4.11) to hold, we need to have

$$\begin{aligned} E[h(\mathbf{U}_1, 1) - h(\mathbf{U}_0, 1)|G = uc, S_1 = 1] &\geq E[h(\mathbf{U}_1, 1) - h(\mathbf{U}_0, 1)|G = t, S_1 = 1] \\ &\geq E[h(\mathbf{U}_1, 1) - h(\mathbf{U}_0, 1)|G = lc, S_1 = 1] \end{aligned} \tag{4.17}$$

A set of sufficient conditions for (4.17) to hold when $\mathbf{U}$ is univariate and the assumptions in §4.2.1 hold is the following: (a) $S_0 = S_1 = 1$ for all units so that all units are in the study population in both periods; (b) $U_1 - U_0$ is independent of U_0 given G; (c) the function

$h(U,1)$ is convex in U so that h has increasing differences in the sense that for u, $u^{'}$, $u^{''}$, $u^{'''}$ such that $u - u^{'} = u^{''} - u^{'''}$ and $u > u^{''}$, the following inequality holds: $h(u,1) - h(u^{'},1) \geq h(u^{''},1) - h(u^{'''},1)$, and (d) $U_1 - U_0|G = lc \preceq U_1 - U_0|G = t \preceq U_1 - U_0|G = uc$. The proof that this set of sufficient conditions implies that (4.17) holds is as follows. Let D_{lc} be a random variable with the distribution of $U_1 - U_0|G = lc$ where D_{lc} is independent of U_0 given G. Then from (c) and (4.5), it follows that

$$E[h(U_0 + D_{lc}, 1) - h(U_0, 1)|G = t] \geq E[h(U_0 + D_{lc}, 1) - h(U_0, 1)|G = lc]. \tag{4.18}$$

Now let D_t be a random variable with the conditional distribution of $U_1 - U_0|G = t$ and D_{uc} be a random variable with the conditional distribution of $U_1 - U_0|G = uc$ where D_t and D_{uc} are independent of U_0 given G. Then from (d) and h being an increasing function, it follows that $E[h(U_0 + D_t)|G = t] \geq E[h(U_0 + D_{lc})|G = t]$. Combining this with (4.18), we have

$$E[h(U_0 + D_t, 1) - h(U_0, 1)|G = t] \geq E[h(U_0 + D_{lc}, 1) - h(U_0, 1)|G = lc]$$

which is equivalent to

$$E[h(U_1, 1) - h(U_0, 1)|G = t] \geq E[h(U_1, 1) - h(U_0, 1)|G = lc]. \tag{4.19}$$

Similarly from (d) and h being an increasing function, it follows that $E[h(U_0 + D_{uc})|G = uc] \geq E[h(U_0 + D_t)|G = uc]$, and from (c) and (4.5), it follows that

$$E[h(U_0 + D_t, 1) - h(U_0, 1)|G = uc] \geq E[h(U_0 + D_t, 1) - h(U_0, 1)|G = t],$$

and combining these, we have that

$$E[h(U_0 + D_{uc}, 1) - h(U_0, 1)|G = uc] \geq E[h(U_0 + D_t, 1) - h(U_0, 1)|G = t]$$

which is equivalent to

$$E[h(U_1,1) - h(U_0,1)|G = uc] \geq E[h(U_1,1) - h(U_0,1)|G = t]. \tag{4.20}$$

Combining (4.19) and (4.20) gives us the desired conclusion.

Proof that (4.9) still holds as long as when (i) in (4.6) holds, (4.11) holds or when (ii) in (4.6) holds, (4.12) holds. When there are time varying confounders, we have that $E[\hat{\beta}_{dd.lc}]$ is the expression on the right hand side of (4.7) plus $E(\epsilon_1|G = t, S_1 = 1) - E(\epsilon_1|G = lc, S_0 = 1)$ and $E[\hat{\beta}_{dd.lc}]$ is the expression on the right hand side of (4.8) plus $E(\epsilon_1|G = t, S_1 = 1) - E(\epsilon_1|G = uc, S_0 = 1)$. When (i) in (4.6) holds, the expression on the right hand side of (4.7) is $\geq \beta$ and the expression on the right hand side of (4.8) is $\leq \beta$. Combining the facts in the last two sentences, we have that if (i) in (4.6) and (4.11) holds, $E[\hat{\beta}_{dd.uc}] \leq \beta \leq E[\hat{\beta}_{dd.lc}]$ and if (ii) in (4.6) and (4.12) holds, $E[\hat{\beta}_{dd.lc}] \leq \beta \leq E[\hat{\beta}_{dd.uc}]$.

4.5.5. Test of Model/Assumptions by Examining the Groups' Relative Trends in the Before Period

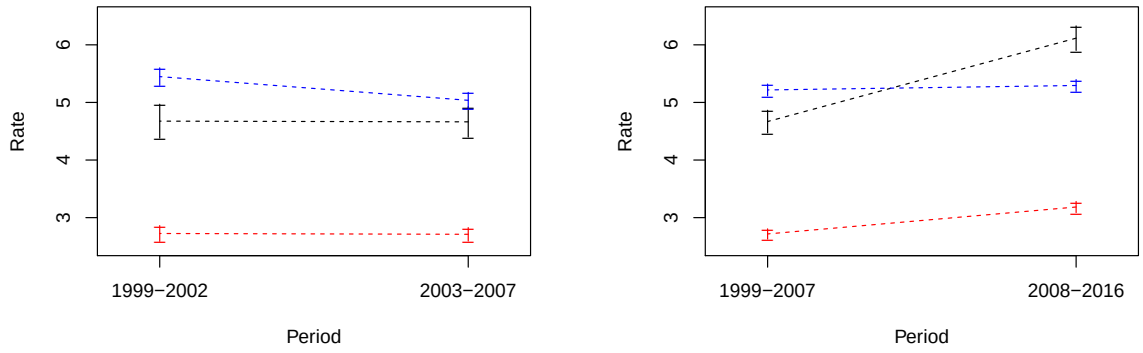
We can test whether the violating pattern (iii) is present in the before period using an intersection-union test (Lehmann, 1952; Berger, 1982), which find evidence (say p-value < 0.05) for (iii) if there is evidence (p-value $< .05$) for both (a) the difference between the upper control group and the counterfactual treated group is larger in the second part of the before period than the first part and (b) the difference between the counterfactual treated group and the lower control group is smaller in the second part than the first part; for the firearm homicide data, splitting the before period into the two parts, 1999-2002 and 2003-2007, (a) gives a p-value of 0.96 and (b) gives a p-value of 0.5, so there is not evidence for (iii) being violated. Pattern (iv) can be tested in a similar way and for the firearm homicide data, there is not evidence for pattern (iv) holding (p-values of 0.04 and 0.5). Ideally, this testing procedure should have sufficient power to reduce the chance of proceeding with the analysis when the assumptions of the model don't, in fact, hold to

an acceptable level. When sample sizes are beyond the control of the investigator or, for example, when dealing with complete counts of firearm homicides where variability depends on the rate itself rather than sampling error, increasing the level of the test can achieve some improvement in power. The p-value is ≥ 0.5 for the test of each alternative, that (iii) holds and that (iv) holds. Hence, α would have to be increased beyond 0.5 to affect the conclusions about the plausibility of our model assumptions.

Alternatively, the presence of violating patterns (iii) and (iv) can be assessed visually without requiring a formal testing procedure. In the left panel of 9 we plot the relative trends of the population-weighted firearm homicide rates for the upper (dashed blue) and lower (dashed red) groups and the counterfactual treated group (dashed black) over the before period. The vertical bars indicate 95% CIs. Visually, there is no strong evidence that pattern (iii) or (iv) is present. The difference between upper controls and counterfactual Missouri and between counterfactual Missouri and the lower controls both get smaller in the latter part of the before period. We can also partially assess whether this pattern might hold over the entire study period, our primary concern, by addressing how the upper and lower control trends compare between the before period and the entire study period. In the right panel of 9 we plot the relative trends of the two control groups and treated group over the entire study period. The dashed black lines are not comparable between panels because the left panel is a counterfactual trend whereas the trend in the right panel is subject to treatment (i.e. permit-to-purchase repeal). However, we can assess the comparability of the pattern of the control group trends between the two panels. They appear similar, with a slight narrowing of the difference in population-weighted firearm homicide rates over time.

When paired with the test described above, visual inspection can answer questions about our model assumptions that our intersection-union tests do not address directly: If we find evidence that pattern (iii) or (iv) is present, are the violations substantial enough to arrest the planned analysis or should we still proceed but with increased caution? If the test doesn't find evidence of a violation is that because our assumptions hold, at least

Figure 9: (Left Panel): Relative trends of the population-weighted firearm homicide rates for the upper (dashed blue) and lower (dashed red) groups and the counterfactual treated group (dashed black) over the before period. The vertical bars indicate 95% CIs. (Right Panel): Relative trends of the population-weighted firearm homicide rates for the upper (dashed blue) and lower (dashed red) groups and the treated group (dashed black) over the entire period. The vertical bars indicate 95% CIs.



approximately, or is it due to large standard errors and/or low power? We recommend that testing and visual inspection should be used in conjunction when assessing the plausibility of the model assumptions.

If one does find evidence for pattern (iii) or (iv) holding in the before period, and if one thinks there has been a structural shift such that the model (4.1)-(4.4) and assumptions (4.5)-(4.6) only start to hold in the latter part of the before period but continue to hold in the after period, one could just use the latter part of the before period. This is similar to the scenario in a difference-in-difference model when there is evidence of a diverging trend during an earlier portion of the pre-intervention period, researchers can restrict the analysis to include only the latter part of the before period with the hope that parallel trend assumption is more likely to be valid (Volpp et al., 2007). However, the finding of pattern (iii) or (iv) in the before period suggests caution.

4.5.6. Analysis Using After Period of 2008-2013

For the period of 2008-2013, Missouri's age-adjusted firearm homicide rate was 5.5, the upper control group's age-adjusted firearm homicide rate was 5.0 and the lower control's age adjusted firearm homicide rate was 2.9. Using an after period of 2008-2013, difference-in-difference estimates for the upper and lower control groups are shown in Table 13. Using an after period of 2008-2013, the interval (4.10) that has a $\geq 95\%$ chance of containing the effect of the repeal on the firearm homicide rate is $[0.2, 1.4]$, corresponding to a 5% to 31% increase in firearm homicides, providing evidence that the repeal increased firearm homicides.

Table 13: Difference-in-difference estimates of effect of repeal of Missouri's permit-to-purchase handgun licensing requirement on firearm homicide rates per 100,000 persons using after period of 2008-2013

Control Group	Estimate	95% CI	% Change Estimate	95% CI
Upper Controls	1.0	[0.6, 1.4]	22%	[14% ,31%]
Lower Controls	0.6	[0.2, 1.0]	17%	[5% ,19%]

4.5.7. Comparison with the Synthetic Control Method

Abadie et al. (2010) proposed constructing a synthetic control group which is a linear combination of multiple control groups that matches the before period outcomes of the treatment group. The synthetic control method provides asymptotically unbiased estimates of the causal effect of treatment assuming that the unmeasured confounders can be represented by a factor model with the factors' effects in each time period being linear with a time-specific slope, whereas our bracketing method only provides bounds under this assumption. However, this assumption is strong and is not generally satisfied in our model (4.1)-(4.4). In the following section we provide a simple example that satisfies the assumptions of our model but for which the estimate returned by the synthetic control method will be biased.

If the types of interaction between history and group in the after period that are of concern have occurred in the before period (e.g., a similar recession occurred in the after period as the before period), then the synthetic control method's matching of the before period outcomes

might enable it to match the treated group's counterfactual trajectory in the after period in the absence of treatment. However, if the types of interaction are different (e.g., there is a more severe recession in the after period or the interactions between poor health and the macroeconomy have been altered by other policy changes), then the synthetic control's matching in the before period does not provide much reassurance unless one has a basis for strong functional form assumptions such as the factors representing the unmeasured confounders' having a linear effect in each time period. In contrast, the bracketing method relies on assumptions such as (4.6) that the unmeasured confounders' effect is increasing (or decreasing) in importance over time over the whole range of the unmeasured confounders that can be assessed using subject matter knowledge without making strong functional form assumptions.

4.5.8. Example of How Synthetic Control Model Assumptions Are Violated in Our Model

For example, suppose U has an exponential distribution in each group with scale 0.2, 0.5 and τ in the lower control, upper control and treated groups respectively where $0.2 < \tau < 0.5$ and $h(U,0) = U$, $h(U,1) = exp(U)$. Then the synthetic control linear combination is $\frac{\tau-0.2}{0.3}\times$ lower control group + $\frac{0.5-\tau}{0.3}\times$ upper control group. For the after period, the linear combination of the mean outcomes for the synthetic control linear combination is $\frac{\tau-0.2}{0.3}\times 1.25+\frac{0.5-\tau}{0.3}\times 2$ while the treated group's counterfactual mean outcome in the absence of treatment is $\frac{-1/\tau}{-1/\tau+1}$, and $\frac{\tau-0.2}{0.3}\times 1.25+\frac{0.5-\tau}{0.3}\times 2 < \frac{-1/\tau}{-1/\tau+1}$ for all $0.2 < \tau < 0.5$. Thus the synthetic control group's after period mean is always less than than the counterfactual after period mean for the treatment group in the absence of treatment.

CHAPTER 5

Estimating Malaria Vaccine Efficacy in the Absence of a Gold Standard Case Definition: Mendelian Factorial Design

Abstract

In this paper we develop methods to identify and estimate malaria vaccine efficacy that do not require a gold-standard case definition of clinical malaria. Instead, we leverage genetic traits that are protective against malaria but not against other childhood illnesses to identify vaccine efficacy in a randomized control trial. The sickle cell trait is one such genetic variant that confers protection specifically against clinical malaria. The method, which we call *Mendelian factorial design*, is inspired by Mendelian randomization studies that use genetic variants as instrumental variables to estimate causal effects of non-randomized exposures. Mendelian factorial design augments a randomized trial with genetic variation to produce a natural factorial experiment, which under realistic assumptions allows for identification of vaccine efficacy. We motivate our methods with a hypothetical study of the pre-erythrocytic vaccine RTS,S where subject-level information on sickle cell status is collected. A robust, covariance adjusted estimation procedure is developed for estimating vaccine efficacy on the risk ratio and incidence ratio scales. Simulations across a number of settings suggest that our estimator has good performance whereas naive methods are systematically biased. We demonstrate that a combined estimator using both our proposed estimator and the standard approach yields significant improvements when the Mendelian factor is only weakly protective. Finally, we extend the Mendelian factorial design framework to time-to-event studies.

5.1. Introduction

In 2017, there were an estimated 219 million cases of malaria of which 92% occurred in Africa and an estimated 435,000 malaria related deaths of which 93% occurred in Africa; nearly every case of malaria in Africa was cause by the parasite *Plasmodium falciparum* (World Health Organization, 2018). Pregnant women and children under the age of 5 are the most vulnerable groups affected by malaria. To date, of more than 30 vaccines under

development, the only vaccine to undergo a pivotal phase III trial is the pre-erythrocytic vaccine RTS,S which has shown to have limited efficacy ($30 - 50\%$ reduction in incidence rates; Mahmoudi and Keshavarz (2017)). Consequently, the continued development of an efficacious P. *falciparum* malaria vaccine has the potential for substantial public health impacts. With so many potential vaccines in the development pipeline, a critically important statistical challenge is to develop methods for estimating vaccine efficacy. To date, accurate estimation of vaccine efficacy against clinical outcomes attributable to P. *falciparum* malaria requires defining reliable case definitions, a task that is made difficult by the unspecific presentation of malaria in endemic areas.

In the absence of a gold standard case definition, efficacy is usually assessed by choosing an inexact case definition which may falsely exclude cases attributable malaria and falsely include non-malaria cases. The established definition of clinical malaria in malaria prevention trials is the presence of a fever with a temperature $\geq 37.5^{o}$C and a P. *falciparum* parasite density above a certain threshold, e.g., 2500 or 5000 parasites per μl of blood (Ter Kuile et al., 2003; The RTS,S Clinical Trials Partnership, 2011; Olotu et al., 2013; Bejon et al., 2013). Case definitions of this form inevitably exclude some true cases and include some false cases due to heterogeneity in immunity and endemicity, fever killing effects, and parasite density measurement error. With specificity < 1, estimates of vaccine efficacy will be biased downward. It has been shown in simulations that such case classification errors have the potential to introduce substantial bias in many settings (Small et al., 2010). Real trial data suggests that these challenges with the standard case definition often go unaddressed. In a multi-site pooled analysis of phase II RTS,S trial data using a fixed 2500 parasite per μl cutoff across all study sites, Bejon et al. (2013) reports estimates of vaccine efficacy against malaria that were as high as 60% in sites with low parasite prevalence and as low as 4% in high parasite prevalence sites. The authors suggest a biological reason for this pattern: RTS,S prevents fevers in only a portion of mosquito bites and in areas where parasite prevalence is high, children are likely bitten more often by infected mosquitos. However this pattern is also consistent with bias due to the fixed case definition having lower

specificity in higher prevalence areas where, due to improved immunity, children can carry higher parasite loads without fever. The standard case definition leaves much unanswered: is the heterogeneity in vaccine efficacy epidemiologically important or just an artifact of bias arising from an inexact case definition?

To overcome the challenges that accompany inexact case definitions, this paper introduces a new method for identifying malaria vaccine efficacy using natural genetic variation, which we call *Mendelian factorial design*. Importantly, the method does not depend on an inexact case definition based on the parasite density, instead using all fevers (or deaths) with any level of parasitemia to estimate efficacy. To identify vaccine efficacy, the new method requires finding and recording genetic variants that provide specific protection against clinical malaria and operate through a different biological pathway than the vaccine being evaluated. A running example in this paper is the sickle cell trait, a hemoglobinopathy that has been shown to protect against malaria at the blood-stage (erythrocytic) of the infection and the RTS,S vaccine, which confers protection against malaria at the pre-erythrocytic, liver-stage of the infection. There are many other vaccine types, e.g., transmission blocking and erythrocytic vaccines, and genetic variants that might satisfy these identifying conditions (Ndila et al., 2018). We will discuss these occasionally throughout the paper and more thoroughly in the discussion.

In the following section, we introduce Mendelian factorial design informally in the context of the more familiar use of genetic variants as instrumental variables in *Mendelian randomization* studies, to which it has many parallels. In §5.3, we present a precise definition of vaccine efficacy and malaria-attributable fevers in a potential outcome framework and propose an identification strategy that holds under a few realistic assumptions. We then provide a simple $\sqrt{n}$-consistent and asymptotically normal covariate-adjusted estimator that is robust to model misspecification and assess its performance in a simulation study over a range of settings. We develop an improved "bounded" estimator that combines the strengths of our estimator with that of the (biased) standard estimator. We demonstrate that it provides

nearly uniform improvement over both estimators. Finally, in §5.4 we extend the Mendelian factorial design to identify and estimate vaccine efficacy in time-to-first event studies.

5.2. Mendelian Factorial Design: Parallels with Mendelian Randomization

When an observed association between a non-randomized exposure and an outcome may be confounded by an unobserved common cause, attributing the association to a causal effect may be misleading (Rosenbaum, 2002c). An instrumental variable (IV) is a covariate that is associated with the exposure but whose only association to the outcome is through a direct pathway to the exposure (Martens et al., 2006). The IV takes the place of the physical randomization in a randomized trial, influencing only the assignment of subjects to exposure or control, setting up a natural experiment and providing an avenue for estimating the causal effect of the non-randomized exposure on the outcome (Rassen et al., 2009). When genetic variants are used as instrumental variables, the corresponding collections of methods are often labeled as the *Mendelian randomization* (MR) approach (Smith and Ebrahim, 2008). For example, Kang et al. (2013) proposed using the hemoglobin S variant (HbS) as an instrument to identify the causal effect of malaria on stunted growth. It is known that heterozygote carriers (HbAS, sickle cell trait) receive protection against clinical malaria whilst those without the variant (HbAA) do not. The first column of Table 14 summarizes the assumptions required of HbAS status such that it can be used in a MR study to identify the causal effect of malaria on stunting. Assumption (1) requires that HbAS status in fact influences exposure; assumption (2) ensures that HbAS status can be treated "as-if" it were randomized; and assumption (3) says that HbAS only effects stunted growth through its influence on exposure. Assumption (4) is required for point identification of the causal effect, although there are other alternative assumptions that can be used to point identify a causal effect such as monotonicity of the effect (Burgess et al., 2017).

Like MR, Mendelian factorial design (MFD) leverages genetic variation to identify a causal effect of interest – in this instance, the efficacy of a vaccine, or roughly, the proportion reduction of disease-attributable outcomes in a population when the vaccine is applied

Table 14: Parallel assumptions for Mendelian randomization and Mendelian factorial design.

Assumption	*Mendelian Randomization*	**Mendelian Factorial Design**
(1)	• HbAS associated with malaria	• HbAS has protective efficacy against malaria-attributable fever
(2)	• No unmeasured confounders associated with HbAS and stunted growth	• HbAS "as-if" randomized
(3)	• Only direct pathway from HbAS to stunted growth is through malaria	• Protection provided by HbAS is specific to malaria-attributable fever
(4)	• HbAS does not modify the effect of malaria on stunting	• No interaction effect between HbAS and vaccine (i.e., independent protective pathways)

(Lachenbruch, 1998). However, instead of addressing the bias that often accompanies a non-randomized exposure, MFD attends to the bias in estimating treatment efficacy arising from an inexact case definition. The natural experiment arising from MR is used for causal inference in the absence of an actual randomized experiment, enabling the investigator to distinguish causal effects from unobserved confounding. The same natural experiment is employed in MFD to augment a two-arm randomized trial and create a simple 2×2 factorial experiment. The factorial design allows the investigator to distinguish efficacy against disease-attributable outcomes even when the case definition incorrectly classifies a material number of non-disease outcomes as cases.

Although MR and MFD address different sources of bias in different settings, their designs share many parallel features. This is illustrated in Table 14, where each assumption of MR in the malaria stunting study corresponds to a closely related assumption that underlies a MFD study of vaccine efficacy against malaria-attributable fever. In more general terms, Assumption (1) says that the *Mendelian gene* or *Mendelian factor*, e.g., HbAS, is relevant. In the MR study this means that it is associated with the non-randomized exposure and in the MFD study this implies that the Mendelian gene is protective against the disease-

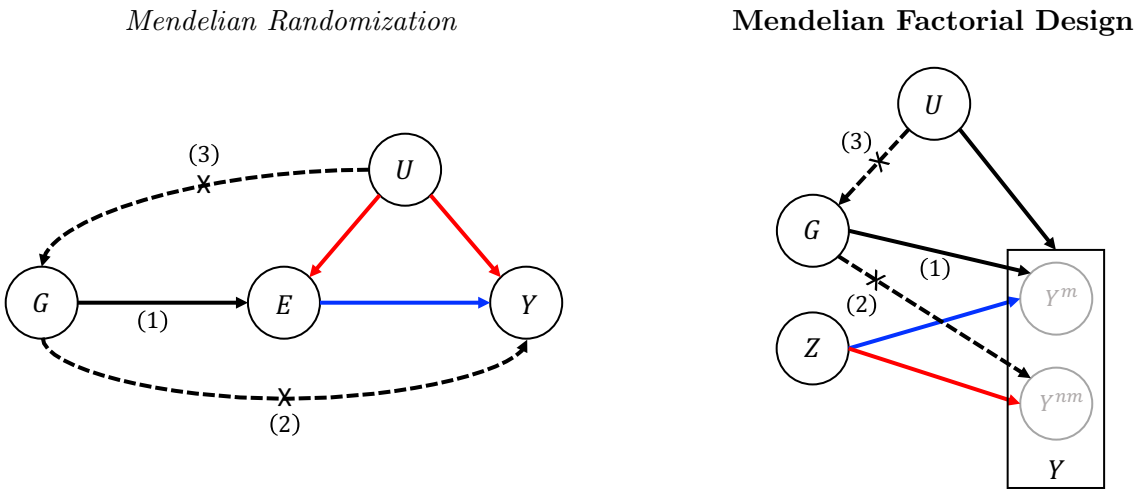


Figure 10: Causal diagrams for MR (left) and MFD (right). Blue arrows indicate the causal quantities of interest and red arrows indicate confounding addressed by each design. Gray variables Y^m and Y^{nm} are unobservable but Y, which doesn't distinguish between them is. Assumptions (1), (2) and (3) correspond to the similarly numbered assumptions in Table 14.

.

attributable outcome. Assumption (2) requires that the Mendelian gene be unconfounded with the outcome of interest or be "as-if" randomized, which implies the former. For the MR study, assumption (3) says that the Mendelian gene has no pleiotropic effects (Smith and Ebrahim, 2008), that is, HbAS only effects stunted growth through its influence on malaria and not through its influence on another modifiable exposure or stunting itself. The parallel assumption for a MFD study is that the Mendelian factor protects against outcomes of any-cause only through reducing disease-attributable outcomes. Finally, assumption (4) says that the causal effect in the MR study does not vary over different levels of the Mendelian gene, or in other words, they do not interact. Similarly, the corresponding assumption in a MFD study is that the vaccine and Mendelian factor do not interact. In other words, the vaccine prevents the same proportion of disease-attributable outcomes at different levels of the Mendelian gene. For example, this assumption is plausible when a malaria vaccine and HbAS provide protection against malaria-attributable fevers through independent biological pathways.

Assumptions (1)-(3) in Table 14 may be better understood encoded in a causal diagram. The causal diagram for MR and MFD are in the left and right panels of Figure 10, respectively. G is a Mendelian gene (or factor), U are unobserved confounders, Z is a randomized vaccine, and E is a non-randomized exposure. Y is the outcome of interest – in the MR study it is stunted growth and in the MFD study it is clinical malaria. Y^m and Y^{nm} are malaria-attributable and non-malaria fevers, respectively. We define these more precisely in §5.3. Finally, the blue arrows indicate the causal quantities of interest and the red arrows indicate the confounding addressed by each design. The causal diagram for MFD is a little bit unusual. Y^m and Y^{nm} are grayed out, emphasizing that we don't observed them, but instead only observe the outcome that is not distinguished by cause, Y. This is represented by the black bounding box. The MFD diagram hints at how the factorial structure and the absence of an arrow between G and Y^{nm} may be used to identify the arrow between Z and Y^m, the vaccine efficacy.

5.3. A Robust Framework for Estimating Vaccine Efficacy: Risk Ratios and Incidence Rate Ratios

5.3.1. Notation: Observed Data

Let $j = 1, \ldots, J$ indicate sites in a multi-site randomized control trial (RCT) and $i = 1, \ldots, I_j$ indicate subjects at each center; then, ij uniquely identifies each subject in the study. We let the total number of subjects in the trial be $n = \sum_j I_j$. For subject ij, let $Y_{ij} \in \{0, 1\}$ be an observed fever or death with any parasitemia; let $Z_{ij} \in \{0, 1\}$ indicate treatment/vaccine status and $G_{ij} \in \{0, 1\}$ indicate sickle cell variant status ($G_{ij} = 1$ if HbAS, $G_{ij} = 0$ if HbAA); and let $X_{ij} \in \mathbb{R}^d$ be a d-dimensional vector of baseline characteristics. Because G is assigned at conception, careful consideration of what variables constitute "baseline" variables should be made. Let D_{ij} and U_{ij} indicate the malaria parasite density in the blood and the level of non-malaria infectious agents, respectively. Define the observed data vector $O_{ij} = (X_{ij}, Z_{ij}, G_{ij}, Y_{ij})$. While U_{ij} is generally not observed in malaria trials, D_{ij} is. However, because the methods developed in this paper do not

require or explicitly model parasite density we omit it from O_{ij}. We drop the subscripts and write $O = (X, Z, G, Y)$ to denote a random draw from a specified population. We will use boldface to indicate corresponding vector and matrix quantities that collect the data of subjects over each site or the entire trial. For example, $\mathbf{Y}_j$ is the $I_j \times 1$ vector that collects all the outcome data for subjects at site j and $\mathbf{Y} = (\mathbf{Y}_1^T, \ldots, \mathbf{Y}_J^T)^T$ is the $n \times 1$ vector that collects the outcome data for all subjects in the multi-site trial. For an example of a matrix quantity, $\mathbf{X}_j$ is an $I_j \times d$ matrix and $\mathbf{X}$ is a $n \times d$ matrix.

Multi-site RCTs, i.e., block randomized, are a common design for vaccine efficacy trials, such as the Phase III RTS,S trial (The RTS,S Clinical Trials Partnership, 2011). Another design that has commonly been employed in studying the protective efficacy of interventions such as insecticide treated bed nets is the clustered RCT (Ter Kuile et al., 2003). The notation is easily adapted for this design, letting j indicate a cluster and $Z_{ij} = Z_j$ for all i, j.

5.3.2. Potential Outcomes and Malaria-Attributable Fever (or Death)

Pitfalls of Standard Case Definitions

The standard case definition of clinical malaria is the presence of a fever ($Y = 1$) and a parasite density above some threshold d ($D > d$). The WHO recommends setting d high enough to achieve specificity $> 80\%$ at all sites to avoid severe underestimation of vaccine efficacy (Moorthy et al., 2007). Sensitivity is also considered when determining d to avoid under-powered studies. The prevailing method used to determine these thresholds is to model risk of fever as a continuous function of observed parasite density among community controls and clinically suspected cases (Smith et al., 1994). However, computing sensitivity and specificity of a case definition requires an estimate of the malaria attributable fraction of fevers (MAFF). Lee and Small (2018) show that obtaining unbiased estimates of MAFF in the presence of measurement error of D and fever killing effects on parasite density is a difficult task, especially when malaria and non-malaria infections can work in conjunction to trigger a fever. Still, when unbiased estimates of MAFF can be obtained, this case

definition will, by design, result in false positive and false negative cases, which may bias corresponding estimates of vaccine efficacy. Heterogeneous immunity and pyrogenic thresholds further complicate estimating treatment efficacy using case definitions that depend on a fixed threshold for D. In practice, standard thresholds without, it seems, careful estimation of specificity and sensitivity such as 2500 parasites per μl (Olotu et al., 2013) and 5000 parasites per μl are commonly used (The RTS,S Clinical Trials Partnership, 2011). As far as we are aware, there are no sufficiently specific case-definitions for death attributable to malaria, the most severe effect of a malaria infection (Moorthy et al., 2007).

Defining Cases Using Potential Outcomes

Potential outcomes are a useful framework to precisely define causal effects of interest (Rubin, 2005). The potential outcomes that we define now are closely related to those defined in Lee and Small (2018). In what follows, we consider Y to indicate the presence of fever but note that the framework we describe is also suitable when Y indicates death. The aforementioned authors motivate their framework with a biological model of malaria and the notion of a pyrogenic threshold (Gravenor and Kwiatkowski, 1998). A pyrogenic threshold can roughly be defined as a level of malaria parasite density above which a fever will be produced. Below the threshold, the infection will not be strong enough to trigger a fever. This threshold may vary from individual to individual based on heterogeneous immunity to symptomatic malaria. In general, we can think of D and U as having thresholds above which a malaria or non-malaria infection, respectively, is strong enough to trigger a fever in the absence of any other infection. We can also consider a curve of threshold pairs for D and U above which a combined infection can trigger a fever.

Because Y indicates the presence of a fever resulting from any infection, we can treat Y as a function of both D and U. Y can be thought of also as a function of Z and G through their effect on D and U. Thus, we can write the potential outcome of Y for treatment z and sickle cell status g as $Y(D(z,g),U(z,g))$ and the observed outcome as $Y = Y(D(Z,G),U(Z,G))$.

$Y(D(z,g),U(z,g))$ factors additively into two natural terms,

$$\underbrace{Y(D(z,g),U(z,g))}_{Y(z,g)} = \underbrace{\{Y(D(z,g),U(z,g)) - Y(0,U(z,g))\}}_{Y^m(z,g)} + \underbrace{Y(0,U(z,g))}_{Y^{nm}(z,g)} . \tag{5.1}$$

Y can be expressed as a potential outcome in terms of d and u or z and g. When it is unambiguous, we may use the abbreviated notation below the "curly" braces in (5.1) to emphasize its dependence on z and g.

The statistical implication of a pyrogenic threshold is that $Y(d,u)$ is monotonic in d for all u. That is, $Y(d,u) \leq Y(d',u)$ for $0 \leq d \leq d'$. This ensures that the first term on the right hand side of (5.1) is non-negative. This term, $Y^m(z,g)$, can be interpreted as *malaria-attributable fever*, or a fever that would not have occurred had a malaria infection been absent. These include cases where D was high enough to trigger a fever in the absence of any other infection and also cases where D was high enough to trigger a fever in conjunction with a non-malaria infection. The second term on the right hand side, $Y^{nm}(z,g)$, is a fever that cannot be attributed to malaria. These are fevers that would have still occurred if malaria parasites were not present. However, when $D > 0$, $Y^m(Z,G)$ is generally not observable because it involves the counterfactual term $Y(0,U(Z,G))$. In the following section we will demonstrate that under a few realistic assumptions we can identify vaccine efficacy without directly observing the fevers (or deaths) attributable to malaria.

A Final Couple Pieces of Notation

With our potential outcome framework in place, we denote the complete data vector by

$$C = (X, Z, G, D, U, \{Y^k(z,g) : k = m, nm;\ z = 0,1;\ g = 0,1\}) .$$

In reality, we only get to observe one potential outcome $Y(Z,G)$ even though all four exist and, as we have argued, Y^{nm} and Y^m cannot be distinguished with certainty so we only observe $O \subset C$. Finally, we may sometimes treat Y, D and U as $p \times 1$ vectors that

correspond to fever status, parasite density, and non-malaria infectious load at p follow-up visits during a RCT.

5.3.3. Vaccine Efficacy: Definitions, Assumptions and Identification

Defining Vaccine Efficacy

We begin this section by defining a general population-level estimand for vaccine efficacy and then outline the assumptions required for identification when we cannot distinguish Y^m from Y^{nm}. We suppose that for each j, C_{ij} (and O_{ij}) are i.i.d. draws from an from an unknown, site-specific target population distribution $P_j \in \mathcal{P}_j$ for all i. We suppose that the overall target population is an equally weighted mixture of the P_j and denote it $P \in \mathcal{P}$. We let $\mu_{zg}(P)$, $\mathcal{P} \to \mathbb{R}$ be a functional of P that depends on treatment and sickle cell status and takes the form $\mu_{zg}(P) = \mathbb{E}_P[f\{Y(z,g)\}]$ for P-measurable functions f that are (1) increasing and (2) linear. We use a superscript m to indicate that the functional is specific to malaria-attributable outcomes and nm to indicate that it is specific to non-malaria outcomes, for instance, $\mu^m_{zg}(P) = \mathbb{E}_P[Y^m(z,g)]$ when $f(y) = y$. Below we give a general definition of efficacy.

Definition 4 (Vaccine/Protective Efficacy)**.** *For a specified function f,*

(i) Vaccine Efficacy is defined as $\tau(g) = 1 - \mu^m_{1g}(P)/\mu^m_{0g}(P)$ for $g = 0, 1$ and

(ii) Protective Efficacy of G is defined as $\nu(z) = 1 - \mu^m_{z1}(P)/\mu^m_{z0}(P)$ for $z = 0, 1$.

We are primarily interested in two simple functions f: $f(y) = y$ and when Y is a $p \times 1$ vector of fevers, $f(y) = \mathbf{1}^T y$. For $f(y) = y$, efficacy as defined in Definition 4 is the proportion reduction in risk of malaria-attributable fever or death were an individual to receive vaccination versus placebo. For $f(y) = \mathbf{1}^T y$, efficacy is defined as the proportion reduction in incidence of malaria-attributable fever were an individual to receive vaccination versus placebo. Since you can only die once, $f(y) = \mathbf{1}^T y$ is not applicable when Y indicates malaria-attributable death. By this same logic, $f(y) = y$ can be used for assessing the risk

of malaria-attributable deaths over arbitrarily long follow-up. However, if we are interested in the risk of developing a malaria-attributable fever over a longer follow-up, during which a subject can have multiple fevers, the function we'd be interested in is $f(y) = \max y$ where $\max y$ is the maximum over the elements of a $p \times 1$ vector y. This function is not a linear function and thus vaccine efficacy against the risk of having at least one malaria-attributable fever over a long follow cannot be handled in this framework. The importance of the linearity of f will become evident shortly.

Identifying Vaccine Efficacy

In §5.2 we informally discussed the assumptions required to identify vaccine efficacy using MFD. Assumptions 1 - 5 formalize the assumptions that are summarized in the second column of Table 14 and provide the basis for the identification of τ proved in Proposition 4.

Assumption 1 (Additivity of μ)**.** *For all $z = 0, 1$ and $g = 0, 1$, $\mu_{zg}(P)$ can be decomposed linearly as*

$$\mu_{zg}(P) = \mu^{m}_{zg}(P) + \mu^{nm}_{gz}(P)\,.$$

Assumption 1 is guaranteed by (5.1) and the linearity of f. We mentioned above that evaluating the vaccine efficacy on the risk scale for malaria-attributable fevers over a long follow-up is problematic when using MFD to estimate τ. The risks of developing a malaria-attributable fever and a non-malaria fever do not satisfy Assumption 1 over long follow-ups because of the non-linearity of $f(y) = \max y$.

Assumption 2 (No Interaction / Independent Protective Pathways)**.** *$\tau(g)$ is constant over $g = 0, 1$ and $\nu(z)$ is constant over $z = 0, 1$. That is, $\tau := \tau(0) = \tau(1)$ and $\nu := \nu(0) = \nu(1)$.*

The biological reasoning behind Assumption 2 is as follows. If a vaccine and genetic trait protect against malaria through independent pathways at different times in the parasite's life cycle, it is plausible that the vaccine will prevent the same fraction of fevers among

those who have the genetic trait ($G = 1$) and those who do not ($G = 0$). Similarly plausible is that possession of the genetic trait will prevent the same fraction of fevers among those to whom the vaccine was administered ($Z = 1$) and to those it was not ($Z = 0$). The goal of pre-erythrocytic vaccines like RTS,S is to provoke an immune response that prevents the parasites from entering the liver, stopping the parasites from ever re-entering the bloodstream and causing clinical symptoms (Regules et al., 2011). In contrast, the sickle cell trait appears to protect against clinical symptoms by inhibiting the growth of parasites once they have re-entered the bloodstream from the liver and by making the host more tolerant to parasite infection (Ferreira et al., 2011; Taylor et al., 2012; Williams, 2011). This suggests that Assumption 2 is satisfied for pre-erythrocytic vaccines and the sickle cell trait.

Assumption 3 (No Interference)**.** *A subject's potential outcomes are functions of its vaccine and sickle cell status alone. That is* $Y_{ij}(\mathbf{z}, \mathbf{g}) = Y_{ij}(z_{ij}, g_{ij})$.

Assumption 3 implies that the treatment and sickle cell status of an individual in the trial effects only their own outcome. This assumption can be rephrased in the context of infectious disease as stating that protection conferred by genetics or vaccine do not materially disrupt disease transmission. This assumption is plausible when considering individuals at two different sites $j \neq j'$ in a multi-site trial but is more complicated for individuals at the same site who may live in close proximity. Stochastic simulation models of malaria vaccination using pre-erythrocytic and blood stage-vaccines have, however, found that transmission effects of such vaccines delivered as they would be in pediatric malaria vaccination programs were minimal (Penny et al., 2008, 2015). Less is known about the suitability of Assumption 3 with respect to the sickle cell variant.

Assumption 4 (Randomization / "as-if" Randomized)**.** *At each site* $j = 1, \ldots, J$*, vaccines are administered randomly and sickle cell status is distributed "as-if" random; that is,*

$$(Z, G) \perp\!\!\!\perp (Y(z, g), Y^m(z, g), Y^{nm}(z, g), X)$$

for all $z, g \in \{0,1\}^2$. *Additionally, we assume that* $Z \perp\!\!\!\perp G$.

The randomization of Z is ensured by the RCT. We assume that within each center, the sickle cell trait is distributed "as-if" random. Population stratification and linkage disequilibrium are common biological violations of the "as-if" random assumption of Mendelian genes (Kang et al., 2013). Briefly, population stratification is when a subgroups that differ on prognostic factors for developing malaria and other childhood illnesses also systematically differ in the prevalence of HbAS. If either the probability of inheriting HbAS or the distribution of prognostic factors is relatively homogenous within a study site, then population stratification would likely not be a material threat to the "as-if" random assumption about G. Linkage disequilibrium is the dependence of gene frequencies at two or more loci (Morton, 2001). If HbAS is in linkage disequilibrium with a gene that affects the risk of childhood illness this may threaten the validity of Assumption 4. Linkage disequilibrium with a gene that affects the risk of malaria is less problematic as this would not violate the exclusion restriction formalized in the next assumption. However, this would require a more delicate treatment of potential outcomes, e.g., we would need to consider the genes in linkage disequilibrium as having potential outcomes depending on G (see VanderWeele and Hernan (2013) for a discussion a related topic of "versions" of treatment).

Assumption 5 (Valid Mendelian Factor)**.** *G is a "valid Mendelian factor" in that it satisfies the following conditions:*

(i) $\nu \neq 0$*; and*

(ii) $1 - \mu_{z1}^{nm}(P)/\mu_{z0}^{nm}(P) = 0$ *for* $z = 0, 1$.

Part (i) of Assumption 5 says that the Mendelian factor is *relevant* to malaria-attributable outcomes. A large body of literature on the protective properties of the sickle cell trait against malaria supports this assumption for HbAS.. Several cohort studies have found evidence that HbAS has 30-50% efficacy against uncomplicated clinical malaria and multiple other case-control and cohort studies of Africa have estimated even greater efficacies against

sever malaria cases of 70-90% (see Gong et al. (2013) for a list of several studies reporting the protective efficacy of HbAS). Assumption 5(ii) can be expressed in terms of potential outcomes by the restriction that U depends only on treatment status, $U(z, g) = U(z)$. The plausibility of Assumption 5(ii) is supported by evidence that the protection conferred by HbAS is "remarkably specific" to malaria, providing little protection to other childhood diseases (Williams et al., 2005).

The following proposition provides a simple, non-parametric identification strategy for treatment efficacy as defined in Definition 4 under Assumptions 1 - 5. We also make the standard assumptions of *consistency*, that $Y = Y(z, g)$ when $Z = z$ and $G = g$, and *positivity*, that the probability of treatment Z and the prevalence of G in each site are both bounded away from 0 and 1.

Proposition 4 (Nonparametric Identification)**.** *Suppose that Assumptions 1 - 5 are satisfied. Then the vaccine efficacy τ is identified from the observed data* $\mathbf{O}$ *as*

$$\tau = 1 - \frac{\mathbb{E}_X[\mathbb{E}_P[f(Y)|\, X, Z = 1, G = 1]] - \mathbb{E}_X[\mathbb{E}_P[f(Y)|\, X, Z = 1, G = 0]]}{\mathbb{E}_X[\mathbb{E}_P[f(Y)|\, X, Z = 0, G = 1]] - \mathbb{E}_X[\mathbb{E}_P[f(Y)|\, X, Z = 0, G = 0]]}, \tag{5.2}$$

where $\mathbb{E}_X$ is the expectation over the marginal distribution of X implied by P.

Proof. The proof of Proposition 4 can be found in Appendix 5.6.1. □

Remarks

Proposition 4 still holds under a weaker version of Assumption 4 requiring only that Z and G are independent of potential outcomes (and of each other) *conditional* on baseline covariates X. Depending on how rich the set of baseline covariates X is, this weaker assumption may be more tenable in the presence of population stratification and linkage disequilibrium.

Equation (5.2) suggests a ratio estimator for τ that is remarkably similar to the Wald estimator used in Mendelian randomization studies (Wald, 1940; Burgess et al., 2017). Without covariates, the Wald estimator of the effect of a non-randomized exposure E on

	$\mathbf{g = 1}$	$\mathbf{g = 0}$
$\mathbf{z = 1}$	$\mu^{nm}(1-\eta) + \mu^{m}(1-\tau)(1-\nu)$	$\mu^{nm}(1-\eta) + \mu^{m}(1-\tau)$
$\mathbf{z = 0}$	$\mu^{nm} + \mu^{m}(1-\nu)$	$\mu^{nm} + \mu^{m}$

Figure 11: 2×2 table for Mendelian factorial design. Each cell represents $\mu_{zg}(P)$ for all combinations of $(z, g) \in \{0, 1\}^2$, which is identified by the observed data $\mathbf{O}$ (Proposition 4). Differencing over the columns then taking the ratio over the rows yields $1 - \tau$. For notational clarity, we drop the subscript from μ_{00}. η is the "spillover efficacy" the vaccine may provide against non-malaria outcomes.

outcome Y using Mendelian gene G can be written as

$$\frac{\bar{Y}_{G=1} - \bar{Y}_{G=0}}{\bar{E}_{G=1} - \bar{E}_{G=0}}, \tag{5.3}$$

where $\bar{V}_{G=g}$ is the sample average of V for individuals with $G = g$. Both (5.2) and (5.3) involve ratios of averages differenced over G. The analogy between (5.2) and the Wald estimator in (5.3) is not by coincidence, but instead arises from a symmetry between the structures of Mendelian randomization with additive effects and Mendelian factorial design with multiplicative effects. More precisely, in Mendelian randomization, the potentially confounded association between the Mendelian gene and the outcome can be decomposed *multiplicatively* into the *additive* effect of the instrument on the exposure and the *additive* effect of the exposure on the outcome. In a Mendelian factorial design, the outcomes can be *additively* decomposed into disease-attributable outcomes and non-disease outcomes upon which the treatment and Mendelian factor have *multiplicative* effects. This simple structure of Mendelian factorial design can be seen clearly in the 2×2 table in Figure 11 of expected outcomes μ_{zg} for different combinations of z and g. Differencing over the columns and taking the ratio over the rows immediately yields $1 - \tau$. The η term in the top row can be thought of as the *spillover efficacy* of the vaccine against non-malaria outcomes. The term drops out when differencing over the columns.

5.3.4. Robust Covariance Adjusted Estimation and Inference

We now propose a simple substitution estimator that allows for covariance adjustment and is robust to arbitrary misspecification of a model for $\mathbb{E}_P[f(Y) \mid X, Z, G]$. The estimation procedure closely resembles that which is developed in Rosenblum and van der Laan (2010) with a couple minor modifications to deal with the factorial structure of our identification procedure and the possibility that HbAS prevalence varies across sites in a multi-site RCT. The simple procedure is detailed below in Algorithm 1 and requires estimating a simple generalized linear model (GLM) with the `R` function `glm` in the `stats` package at most two times. In what follows, we suppose $f(y) = \mathbf{1}^T y$ for expository purposes. That is, we will focus on vaccine efficacy defined as the proportion reduction in the expected number of malaria-attributable fevers.

Algorithm 1 (Estimation of τ)**.** *The following substitution estimator is based on Rosenblum and van der Laan (2010). For clarity, we let* $f(y) = \mathbf{1}^T y$.

---

1. *Estimate* $\mathbb{E}_P[f(Y) \mid X, Z, G]$ *with* `glm` *using a canonical link function (depends on* f*).*
   - *Let the linear part include an intercept, main terms for* Z *and* G*, and interaction* $Z \times G$*, for example,*

     ```
     mu_hat_0 <- glm(f(Y) ~ X*G*Z, family = poisson())
     ```

   - *denote the resulting estimator as* $\hat{\mu}_0(Z, G, X)$.
2. *If there is more than one site and the prevalence of* G *and sample sizes variy across sites, let* `log_mu_hat` $= \log \hat{\mu}_0(Z_{ij}, G_{ij}, X_{ij})$, `p_g` $= (1/I_j) \sum_{i=1}^{I_j} G_{ij}$, `w` $= n/I_j$, *and* `S`

be a categorical variable for site; update $\hat{\mu}_0$ *as follows*

```
mu_hat_1 <- glm(f(Y) ~ offset(log_mu_hat) + S +
                I(w*Z*G/p_g) + I(w*Z*(1-G)/(1-p_g)) + I(w*(1-Z)G/p_g) +
                I(w*(1-Z)*(1-G)/(1-p_g)), family = poisson())
```

and denote the resulting estimator as $\hat{\mu}_1(Z, G, X)$.

3. *Let* $\mu_{zg}(P_n) = \frac{1}{J} \sum_{j=1}^{J} \frac{1}{I_j} \sum_{i=1}^{I_j} \hat{\mu}_1(z, g, X_{ij})$.

4. *Construct the plug-in MFD estimator*

$$\hat{\tau} = 1 - \frac{\mu_{11}(P_n) - \mu_{10}(P_n)}{\mu_{01}(P_n) - \mu_{00}(P_n)}.$$

The estimator $\hat{\tau}$ returned by Algorithm 1 is a special case of a target maximum likelihood estimator (TMLE) (Van der Laan and Rose, 2011). A straightforward modification of Theorem 1 in Rosenblum and van der Laan (2010) yields that $\hat{\tau}$ is consistent and asymptotically normal under mild regularity conditions even when the working model for the conditional expectation is misspecified. Other initial working models $\hat{\mu}_0$ may be used as long as they satisfy certain restrictions on how data-adaptive they are (Van der Laan and Rose, 2011). The weight terms `w` are important because we assume that the target population P is an equally weighted mixture of the site-specific populations P_j but allow for study designs with different different sample sizes across sites. The weights ensure that $\hat{\mu}_1$ solves the efficient influence function estimating equation (see Appendix 5.6.4). Before we state the result, we introduce some some important notation and definitions.

Working Models and their Limits:

Let the maximum likelihood parameter estimates from steps 1 and 2 of Algorithm 1 be $\boldsymbol{\beta}_n^{(0)}$ and $\boldsymbol{\beta}_n^{(1)}$, respectively, and define $\boldsymbol{\beta}_n = [\boldsymbol{\beta}_n^{(0)}, \boldsymbol{\beta}_n^{(1)}]$. Now, recall that P is the true, unknown data generating distribution. We can decompose its corresponding density p as follows as follows

$$p = p(X)p(Z)p_j(G)p(Y \mid X,\, Z,\, G) \tag{5.4}$$

where $p_j(G)$ is the probability a child has the sickle cell trait in site j. $p(Z)$ is assumed to be known, e.g., $p(Z) = 1/2$ in a balanced trial. P_n is our estimate of P where p_n, the density of P_n, can be decomposed similarly as

$$p_n = p_n(X)p(Z)p_{j,n}(G)p_{\beta_n}(Y \mid X,\, Z,\, G) \tag{5.5}$$

where $p_n(X)$ is the empirical distribution of X, $p_{j,n}(G)$ is the observed prevalence of the sickle cell trait among children in enrolled in the study at site j, and $p_{\boldsymbol{\beta}_n}(Y \mid X,\, Z,\, G)$ is the estimated parametric working model for the conditional distribution of Y from steps 1 and 2 in Algorithm 1. We assume that the number of centers are fixed, that they are representative of the population of interest, and that the sites carry equal weight in the population they represent but that the sample sizes I_j might be different. Hence, the observations that make up the empirical distribution are weighted by n/I_j. We assume that the site-level sample sizes I_j grow at the same rate as $n \to \infty$ and so it follows that $p_n(X) \overset{a.s.}{\to} p(X)$ as $n \to \infty$ by the Gilvenko-Cantelli Theorem and an application of the strong law of large numbers gives us that $p_{j,n}(G) \overset{a.s.}{\to} p_j(G)$ for all $j = 1, \ldots, J$ as $n \to \infty$. Let $P_\infty = \lim_n P_n$ and write its density p_∞ as

$$p_\infty = p(X)p(Z)p_j(G)p_{\boldsymbol{\beta}}(Y \mid X,\, Z,\, G) \tag{5.6}$$

where $\boldsymbol{\beta} = [\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}]$ are the maximizers of the expected log-likelihoods of the GLMs in steps 1 and 2 where the expectation is taken over P, if such maximizers exist (see Rosenblum and van der Laan (2010) for further discussion of the existence of $\boldsymbol{\beta}$). When $\boldsymbol{\beta}$ exists, the conditions given in Proposition 5, are sufficient for $\boldsymbol{\beta}_n$ to converge to $\boldsymbol{\beta}$ in probability (Rosenblum and van der Laan, 2009). Note that unless the working parametric model for

the conditional mean is correctly specified, P_∞ will not equal P in general.

For notational convenience, we let $\mathbb{P}$ and $\mathbb{P}_n$ be the expectation operators over P and P_n, respectively. For example, we can write $\mu_{zg}(P_n)$ from step 3 of Algorithm 1 as $\mathbb{P}_n\hat{\mu}_1(z, g, X)$.

Efficient Influence Functions:

For an arbitrary distribution $Q \in \mathcal{P}$, the *efficient influence function* (EIF) for $\mu_{zg}(Q)$ can be written as

$$\varphi_{zg}(Q)(O) = \frac{\mathbb{1}(Z = z)\mathbb{1}(G = g)(f(Y) - \mathbb{E}_Q[f(Y)|\, X,\, Z = z,\, G = g])}{q_j(G = g)q(Z = z)} + \mathbb{E}_Q[f(Y)|\, X,\, Z = z,\, G = g] - \mu_{zg}(Q)\,, \tag{5.7}$$

for $z, g \in \{0, 1\}^2$. We will define $\mu_z(Q) \coloneqq \mu_{z1}(Q) - \mu_{z0}(Q)$ and $\varphi_z(Q) \coloneqq \varphi_{z1}(Q) - \varphi_{z0}(Q)$ for $z = 0, 1$. Standard calculations verify that $\varphi_z(Q)$ is the efficient influence function for $\mu_z(Q)$ by demonstrating that $\varphi_z(Q)$ can be expressed as a pathwise derivative of $\mu_z(Q_\epsilon)$ where Q_ϵ is a parametric submodel of Q such that $Q = Q_{\epsilon=0}$ (Kennedy, 2016). $\varphi_z(Q)$ is said to be a pathwise derivative of $\mu_z(Q_\epsilon)$ if $\mathbb{E}_Q[\varphi_z(Q)S_\epsilon] = \partial\mu_z(Q_\epsilon)/\partial\epsilon|_{\epsilon=0}$ where S_ϵ is the score function of the parametric submodel. Finally, let $\varphi^*_{zg}(Q) = \varphi_{zg}(Q) - \mathbb{E}_P[\varphi_{zg}(Q)\,|\,G]$ and define $\varphi^*_z(Q) \coloneqq \varphi^*_{z1}(Q) - \varphi^*_{z0}(Q)$. We can now state the consistency and asymptotic normality result for $\hat{\tau}$.

Proposition 5 (Consistency and Asymptotic Normality)**.** *In addition to Assumptions 1 - 5, suppose that the number of sites J is fixed, the I_j grow at the same rate as n, and the maximizers $\boldsymbol{\beta} = [\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}]$ exist. Also, suppose that $\|\boldsymbol{\beta}\|_\infty < M$ for pre-specified $M < \infty$. Finally, let (X, Y) be bounded and the terms in the linear parts of the GLMs in steps 1 and 2 of Algoritm 1 be bounded functions on compact subsets of $\{0, 1\}^2 \times \mathbb{R}^d$. Then, $\hat{\tau}$ is consistent and $\sqrt{n}(\hat{\tau} - \tau)$ convergence in distribution to a Gaussian with mean 0 and variance*

$$\sigma^2 = \mathbb{E}_P\left[\frac{\mu_1(P)}{\mu_0(P)^2}(\varphi^*_0(P_\infty)(O) - \mathbb{P}\varphi^*_0(P_\infty)(O)\} - \frac{1}{\mu_0(P)}\{\varphi^*_1(P_\infty)(O) - \mathbb{P}\varphi^*_1(P_\infty)(O)\}\right]^2.$$

When the prevalence of G does not vary between sites and the study is balanced, i.e., $I_j = n/J$ for all j, then step 2 of Algorithm 1 can be skipped. Step 2 may also be skipped if it is a single-site trial. Furthermore, if the working model for $\mathbb{E}_P[f(Y) \mid X, Z, G]$ is correctly specified then σ^2 achieves the semiparametric efficiency bound.

Proof. A sketch of the proof of Proposition 5 can be found in Appendix 5.6.4. □

Remark 3. *We remarked in §5.3.3 that Proposition 4 still holds under a weaker version of Assumption 4 where the independence is conditional on baseline covariates X. Proposition 5 also holds under these weaker assumptions as long as either the working model for the conditional expectation or the conditional assignment models for G and Z are correctly specified.*

Variance Estimator:

Notice that φ_z^* is defined as a projection of φ_z into a lower dimensional subspace and will thus have smaller variance. With that in mind, we can use $\varphi_z(P_n)$ to construct a conservative plug-in estimator of σ^2. In Appendix 5.6.4 we show that P_n solves the EIF estimating equations, that is, $\mathbb{P}_n\varphi_z(P_n)(O) = 0$ for $z = 0, 1$. It follows then that the plug-in estimator of the scale variance σ^2 can be written simply as

$$n \cdot \widehat{\text{var}}(\hat{\tau}) = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{I_j}\left\{\varphi_0(P_n)(O_{ij})\frac{\mu_1(P_n)}{\mu_1(P_n)^2} - \varphi_1(P_n)(O_{ij})\frac{1}{\mu_0(P_n)}\right\}^2 \tag{5.8}$$

With this variance estimator we can now use Proposition 5 to conduct inference on and construct confidence intervals for $\hat{\tau}$.

Naive Estimator:

For simplicity, let's assume that we have balanced sites, i.e., $I_j = n/J$ for all j. Then, had we assumed that all fevers with any parasitemia were malaria-attributable fevers, we might

considered the following *naive estimator*

$$\hat{\tau}_0 = 1 - \frac{\mathbb{P}_n \hat{\mu}_0(1, G, X)}{\mathbb{P}_n \hat{\mu}_0(0, G, X)}. \tag{5.9}$$

The naive estimator corresponds with standard estimates of VE with respect to a commonly used secondary case definition of the presence of a fever ($Y = 1$) and any positive parasite density ($D > 0$) (Olotu et al., 2013). We will see in the following section that the estimator is systematically biased but can be combined with our MFD estimator $\hat{\tau}$ to construct an estimator that outperforms both estimators on their own.

5.3.5. Simulation Study: Comparison to Naive Identification Strategy

In this section we investigate the performance of our proposed MFD estimator $\hat{\tau}$ and compare it to that of the naive estimator $\hat{\tau}_0$ that assume all fevers with any parasitemia are malaria fevers. As in the previous section, we consider $f(y) = \mathbf{1}^T y$ and thus τ is the proportion reduction in expected number of malaria-attributable fevers. We consider a single-site RCT $J = 1$ with equal sized vaccine and placebo arms and the prevalence of HbAS set to 20% based on existing estimates of the prevalence in sub-Saharan Africa (Ter Kuile et al., 2003; Elguero et al., 2015).

We simulate the number of malaria-attributable fevers and non-malaria fevers from negative binomial distributions in a single year of follow up for each individual. Evidence suggests that the negative binomial distribution fits the empirical distribution of the number of clinical malaria events an individual experiences well (Olotu et al., 2013). We can also see from (5.1) that Y^m and Y^{nm} are negatively dependent – a fever cannot be both attributable to malaria and have been present in the absence of malaria. To model this dependence structure, we use a Gaussian copula with negative dependence parameter $\rho = -0.1$ (Genest and Neslehová, 2007). The negative dependence is modest when fevers are rare but will be more pronounced in areas where fever risk is higher, e.g., villages with poor sanitation. We suppose there is a single observed covariate X that enters the conditional mean function

for both $\mathbf{1}^T Y^m$ and $\mathbf{1}^T Y^{nm}$ and that there is unobserved heterogeneity in the conditional means between individuals. The generating distributions were calibrated so that the average number of any-cause fevers is 1.5 per child-year. We also calibrated the generating distributions to achieve different levels of case specificity, which we define formally below.

Two different sample sizes were chosen to assess how performance improved asymptotically: 1000 subjects per trial arm, i.e., $n = 2000$, and 2500 subjects per trial arm, i.e., $n = 5000$. For $n = 2000$, we consider $5 \times 2 \times 2$ different combinations of vaccine efficacy $\tau \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$, protective efficacy of the Mendelian factor $\nu \in \{0.3, 0.5\}$, and specificities $s \in \{0.5, 0.8\}$. Formally, we define specificity s here as the expected number of malaria-attributable fevers divided by the expected number of fevers with any parasitemia and of any cause under placebo,

$$s = \frac{\mathbb{E}_P\left[\mathbf{1}^T Y^m(0, G)\right]}{\mathbb{E}_P[\mathbf{1}^T Y(0, G)]} = \frac{p(G=1)\mu_{01}^m + p(G=0)\mu_{00}^m}{p(G=1)\mu_{01} + p(G=0)\mu_{00}}\,. \tag{5.10}$$

The specificity choices are motivated by Mabunda et al. (2009), which estimated the specificity of standard case definitions using > 0 and > 2500 parasites per μl cutoffs for children under five years of age to be roughly 50% and 80%, respectively. The strength of the Mendelian factor was calibrated to the aforementioned estimates of the protective efficacy of HbAS. We consider the same combinations for $n = 5000$ except only for the weaker protective efficacy setting. The spillover efficacy η was assumed to be zero. Finally, we allow for non-constant vaccine and Mendelian protective efficacy – both τ_i and ν_i are log-normally distributed with mean τ and ν, respectively, and standard deviation approximately 0.05 giving coefficients of variation of about 8-17%. Rather than conferring complete protection to a fraction of the vaccinated subjects, all vaccinated subjects receive partial protection. This is consistent with evidence that RTS,S is a "leaky" vaccine, providing at least partial protection to all recipients of the vaccine (Moorthy and Ballou, 2009). Each setting was simulated $N_{sim} = 5000$ times. More details of the simulation settings can be found in

Appendix 5.6.5.

We use Poisson regression for the initial working model estimate in Algorithm 1 and because $J = 1$, we skip step 2. The results of the simulation study are summarized in the following two tables. In Table 15 we compare the absolute proportional bias and root mean squared error (RMSE) of the $\hat{\tau}$ (MFD) and $\hat{\tau}_0$ (Naive). For $n = 2000$, the MFD estimator has very good bias properties, with proportional bias less than 5% for most settings. The RMSE increases for Mendelian genes that are less protective and as the specificity decreases. The only area of poor bias performance is when the Mendelian gene is weakly protective ($\nu = 0.3$) and the specificity is low (0.5). However, although the estimator in this setting has high variance the bias and power at $\tau = 0.7$ are reasonably adequate. Like the presence of small sample bias and high variance for weak instruments in instrumental variable analysis (Imbens and Rosenbaum, 2005), the poor performance appears to be a small sample property as the performance for the weak Mendelian gene, low specificity setting improves notably for $n = 5000$.

The bias in the naive estimator only varies over the different specificity settings and does not improve as the sample size grows. In fact, because there is no spillover efficacy, the proportional absolute bias is equal to $1 - s$. You can see from the table that the RMSE is driven almost entirely by the bias component for the naive estimator and it actually increases in absolute terms as the vaccine efficacy increases. These properties lead to very poor coverage of confidence intervals derived from the naive estimator. That one should expect efficacy estimates to be biased by as much as 20% when specificity is at the level recommended by the WHO should be cause for concern.

Table 16 compares the coverage of two-sided 95% confidence intervals and the power against the two-sided alternative at 5% significance for the MFD and naive procedures. The MFD confidence interval has correct or conservative coverage and decent power in even some of the more unfavorable settings. Even in the small sample, weakly protective Mendelian gene, and low specificity setting the power is not negligible at higher vaccine efficacies. Because

Table 15: Proportional absolute bias and root mean squared error (RMSE) of MFD and naive estimators using $N_{sim} = 5000$ simulations.

		Specificity = 0.8				Specificity = 0.5			
		Prop. |Bias|		RMSE		Prop. |Bias|		RMSE	
ν	τ	MFD	Naive	MFD	Naive	MFD	Naive	MFD	Naive
n = 2000									
0.50	0.30	0.03	0.20	0.15	0.07	0.12	0.50	0.29	0.15
	0.40	0.02	0.20	0.14	0.08	0.07	0.50	0.27	0.20
	0.50	0.02	0.20	0.12	0.10	0.05	0.50	0.25	0.25
	0.60	0.02	0.20	0.11	0.12	0.04	0.50	0.24	0.30
	0.70	0.01	0.20	0.10	0.14	0.02	0.50	0.22	0.35
0.30	0.30	0.12	0.20	0.36	0.07	0.59	0.50	5.47	0.15
	0.40	0.11	0.20	0.30	0.09	0.32	0.50	4.54	0.20
	0.50	0.08	0.20	0.29	0.10	0.21	0.50	5.80	0.25
	0.60	0.06	0.20	0.24	0.12	0.18	0.50	4.61	0.30
	0.70	0.03	0.20	0.21	0.14	0.11	0.50	1.69	0.35
n = 5000									
0.30	0.30	0.06	0.20	0.18	0.06	0.18	0.50	0.37	0.15
	0.40	0.04	0.20	0.17	0.08	0.13	0.50	0.34	0.20
	0.50	0.02	0.20	0.14	0.10	0.08	0.50	0.30	0.25
	0.60	0.02	0.20	0.13	0.12	0.06	0.50	0.29	0.30
	0.70	0.01	0.20	0.12	0.14	0.04	0.50	0.27	0.35

Table 16: Coverage of two-sided 95% confidence interval and power against two-sided alternative at 5% significance level of MFD and naive estimators using $N_{sim} = 5000$ simulations.

		Specificity = 0.8				Specificity = 0.5			
		Coverage		Power		Coverage		Power	
ν	τ	MFD	Naive	MFD	Naive	MFD	Naive	MFD	Naive
n = 2000									
0.50	0.30	0.95	0.50	0.56	1.00	0.96	0.00	0.29	0.99
	0.40	0.95	0.17	0.79	1.00	0.96	0.00	0.44	1.00
	0.50	0.96	0.01	0.94	1.00	0.96	0.00	0.59	1.00
	0.60	0.96	0.00	0.99	1.00	0.96	0.00	0.73	1.00
	0.70	0.96	0.00	1.00	1.00	0.97	0.00	0.84	1.00
0.30	0.30	0.95	0.50	0.30	1.00	0.95	0.00	0.19	0.99
	0.40	0.95	0.16	0.44	1.00	0.96	0.00	0.25	1.00
	0.50	0.96	0.01	0.57	1.00	0.96	0.00	0.33	1.00
	0.60	0.96	0.00	0.74	1.00	0.97	0.00	0.42	1.00
	0.70	0.96	0.00	0.86	1.00	0.98	0.00	0.51	1.00
n = 5000									
0.30	0.30	0.95	0.11	0.45	1.00	0.95	0.00	0.26	1.00
	0.40	0.96	0.00	0.68	1.00	0.96	0.00	0.38	1.00
	0.50	0.95	0.00	0.87	1.00	0.96	0.00	0.52	1.00
	0.60	0.96	0.00	0.96	1.00	0.97	0.00	0.66	1.00
	0.70	0.96	0.00	0.99	1.00	0.97	0.00	0.77	1.00

the naive estimator is systematically biased, the coverage properties are extremely poor in all settings. The variance of the naive estimator tends to be far smaller than the MFD estimator. This fact coupled with the systematic bias leads to high power and low coverage across all settings.

Even in settings where the mean proportional bias is high, the MFD estimator appears to have the desirable property of being median unbiased. Figure 12 demonstrates this for $\tau = 0.3, 0.5$, and 0.7 in all six settings detailed in Tables 15 and 16. The dashed lines indicates the true value of τ, the boxes represents the IQRs, and the whiskers indicate the

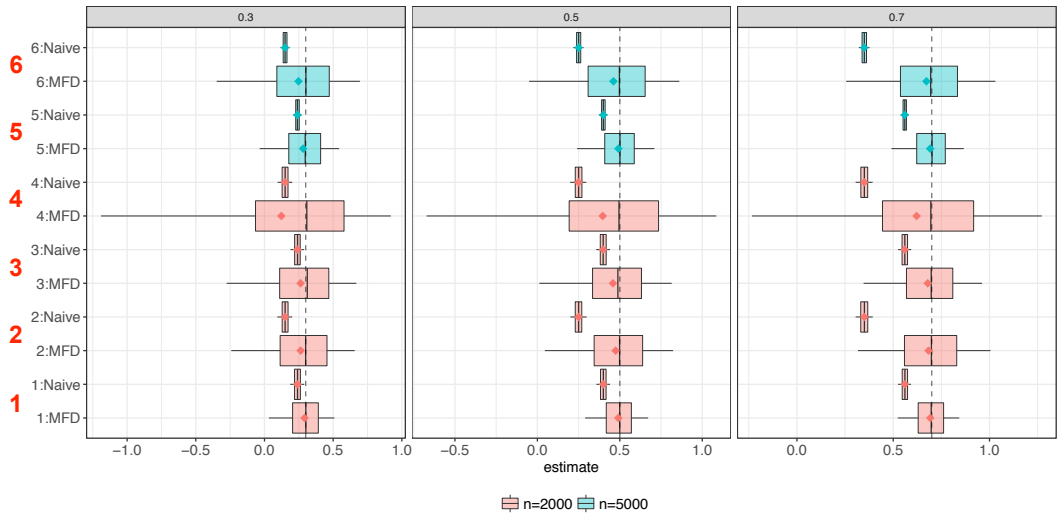


Figure 12: Distributions of simulated MFD estimator $\hat{\tau}$ and naive estimator $\hat{\tau}_0$ over several settings and $\tau = 0.3, 0.5, 0.7$. Setting **1**: strong factor ($\nu = 0.5$), high specificity ($s = 0.8$); setting **2**: strong factor, low specificity ($s = 0.5$); settings **3** and **5**: weak factor ($\nu = 0.3$), high specificity; settings **4** and **6**: weak factor, low specificity. Dashed lines indicate true efficacy, boxes indicate IQRs, whiskers are 5% and 95% quantiles, diamonds are mean estimate, and vertical solid lines are median estimates.

5% and 95% quantiles of the simulated estimates. The diamonds indicate the means and the vertical solid lines the medians. The worsening performance of the MFD as the Mendelian gene weakens is clear (settings 2 vs. 1, 4 vs. 3, and 6 vs. 5) as is the improvement in the weak Mendelian gene settings as the sample size grows (settings 5 vs. 3 and 6 vs. 4). The naive estimates have much lower variance but are systematically mean and median biased downward. The MFD estimates are also mean biased downward. Although the estimator is asymptotically normal, the ratio of means that are jointly asymptotically normal may have a peculiar non-normal form in finite samples which may explain this particular pattern of bias (Marsaglia et al., 2006).

A Simple Correction to the Naive Estimator

The observation that the proportional absolute bias of $\hat{\tau}_0$ is equal to $1 - s$ in Table 15 for all settings comes from the fact that, when the sample sizes are balanced across sites, an immediate application of Theorem 1 of Rosenblum and van der Laan (2010) yields that $\hat{\tau}_0$ is a consistent estimator of $s\tau + (1-s)\eta$. When samples are small or the Mendelian gene is only weakly protective, we observed that the MFD estimator will be less powerful and may suffer from small sample bias. In such settings, the asymptotic limit of $\hat{\tau}_0$ suggests a simple correction to the naive estimator: dividing by s. Call this the *s-corrected estimator*, which is consistent for τ when $\eta = 0$ and s is known or consistently estimated itself. If instead we only have a $1-\beta$ confidence interval for s, $\mathcal{C}_\beta$, then we can still construct a valid confidence interval for τ using the s-corrected estimator. Let $\text{CI}_{s,\alpha+\beta}$ be a $1 - \alpha - \beta$ confidence interval for τ constructed using the s-corrected estimator and define the s-corrected $1-\alpha$ confidence interval $\text{CI}_{corr,\alpha} = \bigcup_{s \in \mathcal{C}_\beta} \text{CI}_{s,\alpha+\beta}$. Berger and Boos (1994) show that one can construct valid p-values by maximizing a non-pivotal p-value over the confidence set of a nuisance parameter. This result is easily inverted, providing a procedure to construct valid confidence intervals from which it follows that $\text{CI}_{corr,\alpha}$ will have the correct (conservative) coverage for τ. However, as we mentioned earlier in §5.3.2, estimating and conducting inference about s is challenging. However, we can still make use of the naive estimator even when we do not know s.

The Best of Both Worlds? Combining MFD and Naive Estimators

The naive estimator and the MFD estimator have complementary strengths and different weaknesses – $\hat{\tau}_0$ tends to be more efficient but is asymptotically biased and $\hat{\tau}$ is consistent but has higher variance and requires larger sample sizes. The naive estimator also has the nice feature of being a consistent lower bound of τ as long as $\eta \leq \tau$ – it is very plausible that a well designed vaccine will have a higher vaccine efficacy than spillover efficacy. Additionally,

we have the logical constraint that $\tau \leq 1$ since a treatment efficacy greater than one would imply that it would be possible to have a negative number of malaria-attributable fevers, a scenario eliminated by the assumption that $Y(d,u)$ is monotonically increasing in d for all levels u. The following proposition constructs an estimator that uses these upper and lower bounds to improve the performance of $\hat{\tau}$ in difficult settings, e.g., small samples, weak Mendelian factor, and low specificity.

Proposition 6 (Bounded Estimator)**.** *Suppose that $\eta \leq \tau$. Let $\hat{\tau}_0$ be the naive estimator and let*

$$L_\alpha = \hat{\tau} - \Phi(1-\alpha)\widehat{\text{var}}(\hat{\tau})^{1/2}$$
$$L_{0,\alpha} = \hat{\tau}_0 - \Phi(1-\alpha)\widehat{\text{var}}(\hat{\tau}_0)^{1/2}.$$

Define the upper confidence bounds U_α and $U_{0,\alpha}$ similarly. Then (i)

$$\hat{\tau}_{bnd} = \min\left[1, \max\left\{\hat{\tau}, L_{0,\tilde{\alpha}}\right\}\right] \tag{5.11}$$

is consistent for τ when $\tilde{\alpha}$ is bounded away from 1; and (ii), for $0 \leq \alpha_0 \leq \alpha/2$

$$CI_{bnd,\alpha} = \left[\max\{L_{\alpha/2-\alpha_0}, L_{0,\alpha_0}\}, \min\left\{1, U_{\alpha/2}\right\}\right] \tag{5.12}$$

is an asymptotically valid $1-\alpha$ confidence interval for τ.

Proof. The proof of Proposition 6 can be found in Appendix 5.6.3. □

Because the naive estimator has much smaller variance than the MFD estimator, one will likely choose α_0 and $\tilde{\alpha}$ to be much smaller than α. In general, it makes sense to choose $\tilde{\alpha} = \alpha_0$ to ensure that $\hat{\tau}_{bnd} \in \text{CI}_{bnd,\alpha}$.

The lower bound used to construct $\hat{\tau}_{bnd}$ acts like a high probability, stochastic lower bound

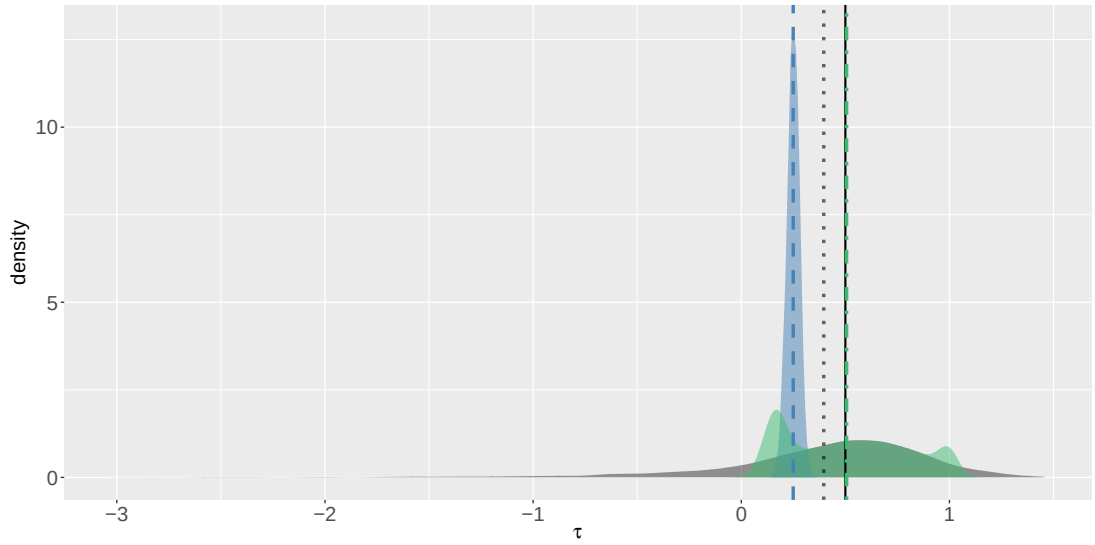


Figure 13: Densities and means of naive estimator (blue, dash) MFD estimator (gray, dot), and bounded estimator (green, dot-dash); True vaccine efficacy (black, solid). For the setting with $n = 1000(\times 2)$, $\tau = 0.5$, $\nu = 0.3$, spec. $= 0.5$, and $\tilde{\alpha} = 0.001$

to τ while 1 is an exact upper bound. When the Mendelian gene is weak, the denominator in $\hat{\tau}$ can sometimes be very close to zero, leading to unreliable estimates of vaccine efficacy. The bounded estimator is designed to mitigate the effect of these cases while keeping the median unbiasedness of $\hat{\tau}$ intact. Figure 13, illustrates how this plays out in a the setting where the $\hat{\tau}$ performs poorly with $n = 2000$, a weak Mendelian factor, low specificity, and $\tau = 0.5$. The figure shows the estimated densities and means of the naive estimator (blue, dash), MFD estimator (gray, dot), and bounded estimator (green, dot-dash) over 5000 simulations. The true vaccine efficacy is indicated by the black solid line. As expected, $\hat{\tau}$ is much more variable and less biased than $\hat{\tau}_0$, but still materially biased in this setting. The bounded estimator $\hat{\tau}_{bnd}$ is nearly mean unbiased. You can see how it achieves this by "clumping" unreliable MFD estimates near the stochastic lower bound and the exact upper bound.

We also assess how well $\hat{\tau}_{bnd}$ performs in the same setting as in Figure 13 but with an even smaller sample size ($n = 1000$). Using $\alpha = 0.05$, $\alpha_0 = 0.001$, and $\tilde{\alpha} = 0.001$, we find that $\hat{\tau}_{bnd}$ has substantially improved bias and RMSE while $\text{CI}_{bnd,\alpha}$ has very favorable

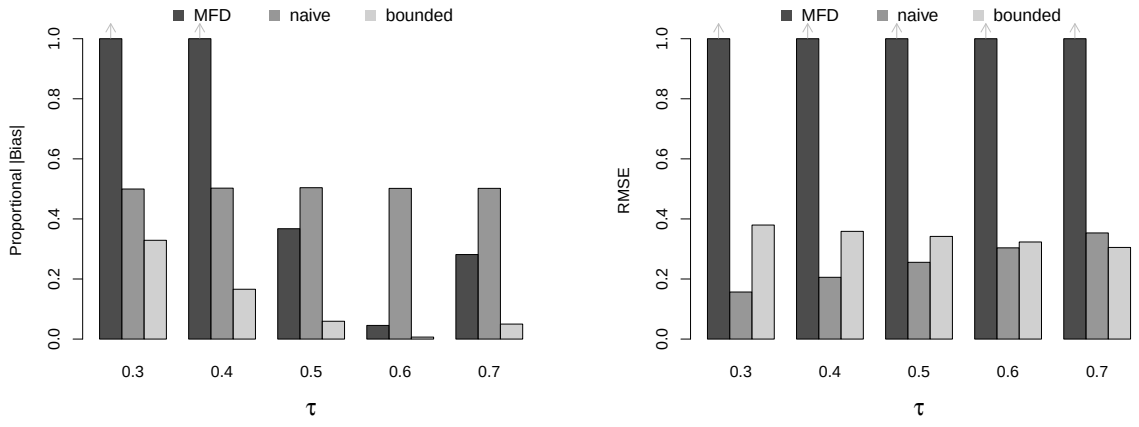


Figure 14: Absolute proportional bias (left panel) and RMSE (right panel) for MFD, naive, and bounded estimators with sample size $n = 1000$ and $N_{sim} = 5000$ simulations. For bias and RMSE values above 1, only a maximum of 1 is shown. The actual absolute proportional bias values for the MFD estimator are 2.21 and 1.56 for $\tau = 0.3$ and 0.4, respectively. The actual RMSE values for the MFD estimator from left to right are $18.58, 67.90, 32.29, 8.42,$ and 6.75.

power while maintaining the correct (conservative) coverage. In Figure 14, we compare the absolute proportional bias (left panel) and RMSE (right panel) of the three estimators. Absolute proportional bias and RMSE values larger than 1 are not shown. The bounded estimator uniformly outperforms both the MFD and naive estimators in terms of bias and has performance comparable to that of the MFD estimator estimated on a sample five times as large ($n = 5000$) for $\tau = 0.5, 0.6,$ and 0.7. The RMSE of $\hat{\tau}_{bnd}$ shows large improvements over the $\hat{\tau}$ and is comparable to the RMSE of $\hat{\tau}$ estimated in the same setting on a sample size of $n = 5000$ for all values of τ considered.

The power and coverage properties of $\text{CI}_{bnd,\alpha}$ give the clearest picture of how the complementary strengths of the MFD and naive procedures are retained by the bounded procedure. In the left panel of Figure 15, we see that $\text{CI}_{bnd,0.05}$ has only marginally more conservative coverage than the MFD confidence interval $[L_{0.025}, U_{0.025}]$ (right panel) while retaining the favorable power properties of the naive confidence interval $[L_{0,0.025}, U_{0,0.025}]$ (left panel). The dashed line in the right panel indicates 95% coverage.

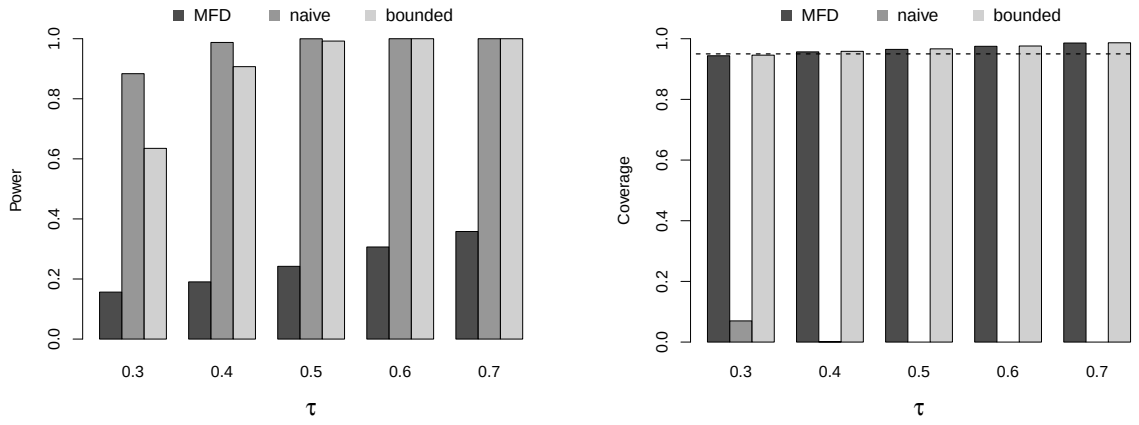


Figure 15: Power against two-sided alternative at 5% significance (left panel) and coverage of 95% confidence interval (right panel) for MFD, naive, and bounded estimators with sample size $n = 1000$ and $N_{sim} = 5000$ simulations. The dashed black line in the right panel indicates 95% coverage.

Importantly, the superior performance of the bounded estimator in this particular setting does not appear to come at the expense of performance in the settings where $\hat{\tau}$ generally does well, improving on the $\hat{\tau}$ in almost all settings investigated in the simulation study.

5.4. Time-to-First Malaria Fever: Mendelian Factorial Design Under The Proportional Hazards Assumption

The World Health Organization (WHO) recommends that the primary endpoint in pivotal Phase III trials assessing the efficacy of malaria vaccines be the time-to-first malaria fever and that the efficacy be measured as one minus the hazard ratio returned by a Cox proportional hazard regression (Moorthy et al., 2007). In this section we use our potential outcome framework to precisely define vaccine efficacy in time-to-first malaria fever studies and show how it can be identified using MFD and estimated using data on time-to-first fever of any cause.

5.4.1. Parameter Identification and Estimation under Modeling Assumptions

In this section we can think about our clinical outcomes as stochastic processes indexed by t, $Y := \{Y_t\}_{t \geq 0}$. This allows us to consider quantities like the time to first fever, $T = \min\{t : Y_t = 1\}$.

If we follow the WHO recommendation above in the context of the potential outcome framework developed in §5.3.2, a natural quantity to study in a time-to-event analysis of vaccine efficacy is the hazard rate of malaria-attributable fevers, i.e., the instantaneous risk of developing a malaria-attributable fever at time t conditional on being free of malaria-attributable fevers up to time t. The risk set used in this hazard function includes individuals who have experienced non-malaria fevers prior to time t. This is analogous to a *subdistribution hazard* in the competing risks literature where non-malaria fevers can be viewed as a competing risk (Fine and Gray, 1999). Unfortunately, when Y^m is not observable, vaccine efficacy based on a hazard ratio using the definition of malaria-attributable fever in (5.1) is not identifiable using MFD. This is due to the fact that on the hazard ratio scale, Assumption 1 (additivity) does not hold when Y^m and Y^{nm} are dependent – recall, if Y_t^{nm} equals 1 then Y_t^m must equal 0 but is otherwise free to take values in $\{0, 1\}$. Only under independence (or independence conditional on X) can we additively decompose the hazard of developing a fever of any cause into the hazard of developing a malaria-attributable fever and the hazard of developing a non-malaria fever. That said, under a few additional assumptions we can still identify vaccine efficacy in terms of the hazard rate for malaria-attributable fevers in the *absence of non-malaria infections*. This hazard is analogous to a cause-specific hazard in the competing risks literature when competing risks can be considered conditionally independent (Hsu et al., 2017).

Malaria-attributable Fevers in the Absence of Competing Infections

If we assume that there is no combined effect of malaria and non-malaria infections on fevers (Assumption 2(iii) of Lee and Small (2018)), then we can provide an alternative

decomposition of Y to the decomposition presented in (5.1):

$$Y(D(z,g),U(z)) = Y(D(z,g),0) \vee Y(0,U(z))\,, \tag{5.13}$$

where $Y(D(z,g),0)$ are malaria-attributable fevers in the absence of non-malaria infections, $Y(0,U(z))$ are non-malaria fevers in the absence of malaria infections, and $\vee$ is a pairwise maximum. For stochastic processes, $A \vee B = \{\max(A_t, B_t)\}_{t\geq 0}$. We will refer to $Y(D(z,g),0)$ and $Y(0,U(z))$ as *isolated* malaria fevers and *isolated* non-malaria fevers. If X sufficiently captures shared risk factors for malaria and non-malaria infections, then it is plausible that $D(z,g) \perp\!\!\!\perp U(z) \mid X$ and thus $Y(D(z,g),0) \perp\!\!\!\perp Y(0,U(z)) \mid X$ for all z, g.

One might wonder if isolated malaria fevers are the endpoint of greatest interest. Perhaps malaria-attributable fevers as defined in (5.1) are more representative of the real-world burden of malaria infections. Regardless, one could argue that the standard case definition of clinical malaria, e.g., $Y = 1$ and $D > 2500$ parasites per μl, is an approximation of an isolated malaria fever. This argument has two parts: (1) the fixed threshold suggests that the standard case definition does not consider the possibility of combined effects of malaria and non-malaria fevers; and (2) the standard case definition does not consider whether a non-malaria fever would have occurred had there been no malaria infection. The approximation is rough, however, because while the definition of isolated malaria fever allows for individual-specific pyrogenic thresholds, the standard case definition does not.

Identifying Vaccine Efficacy

We can now define the potential time to first isolated malaria fever as $T^m(z,g) = \min\{t : Y_t(D(z,g),0) = 1\}$ and the potential time to first isolated non-malaria fever as $T^{nm}(z,g) = \min\{t : Y_t(0,U(z)) = 1\}$. We can write the conditional hazard functions for $T^m(z,g)$ and $T^{nm}(z,g)$ as

$$\lambda^k_{zg}(t \,|\, X) = \lim_{\Delta \to 0_+} \mathbb{P}\Big(T^k(z,g) \in [t, t+\Delta) \,|\, T^k(z,g) \geq t,\, X\Big)/\Delta \tag{5.14}$$

for $k = m, nm$ and all z, g. We drop the superscript k to indicate the conditional hazard function for the a fever of any cause and define

$$T(z,g) = \min\{t :\, Y_t(z,g) = 1\} = \min\{T^m, T^{nm}\}\,.$$

Assumption 6 (Proportional Hazards)**.** *$\lambda^k_{zg}(t \,|\, X)$, $k = m, nm$ follow a Cox proportional hazard model with baseline hazard functions $\lambda^k(t)$, $k = m, nm$. That is,*

(i) $\lambda^m_{zg}(t \,|\, X) = \lambda^m(t) \exp\{\log\kappa + \log(1-\tau)z + \log(1-\nu)g + \boldsymbol{\beta}_m^T X\}$, *and*

(ii) $\lambda^{nm}_{zg}(t \,|\, X) = \lambda^{nm}(t) \exp\{\log\phi + \log(1-\eta)z + \boldsymbol{\beta}_{nm}^T X\}$

where $\nu > 0$.

In equation (i) above, vaccine efficacy τ is equal to one minus the hazard ratio of isolated malaria fever under vaccination versus placebo. In (ii), η can again be thought of as a "spillover efficacy" term. Note that Assumptions 2 and 5 are satisfied under these modeling assumptions.

Assumption 7 (Conditionally Independent First Fever Processes)**.** *The time to first isolated malaria fever and time to first isolated non-malaria fever are conditionally independent. That is,*

$$T^m(z,g) \perp\!\!\!\perp T^{nm}(z) \mid X \quad \textit{for } (z,g) \in \{0,1\}^2$$

Because $T^m(z,g)$ depends only on $Y(D(z,g), 0)$ and $T^{nm}(z,g)$ depends only on $Y(0, U(z))$, Assumption 7 is satisfied when $Y(D(z,g), 0) \perp\!\!\!\perp Y(0, U(z)) \mid X$ for all z, g. We previously argued that this condition is plausible when X sufficiently describes shared risk factors for malaria and non-malaria infections.

Assumption 8 (Shared Conditional Baseline Hazard Function)**.** *Time-to-first isolated*

malaria fever and time to first isolated non-malaria fever have a shared baseline hazard function conditional on X. That is, $\lambda(t) \coloneqq \lambda^m(t) = \lambda^{nm}(t)$ for all $t \geq 0$ and $\boldsymbol{\beta} \coloneqq \boldsymbol{\beta}_m = \boldsymbol{\beta}_{nm}$.

Under these additional assumptions, we can use an MFD strategy to identify τ in time-to-first fever studies.

Proposition 7 (Identification of τ)**.** *Suppose that Assumptions 6 - 8 hold. Then the hazard function for $T(z, g)$ can be written as*

$$\lambda_{zg}(t \,|\, X) = \lambda(t) \exp\{\alpha + \omega z + \gamma g + \lambda z \times g + \boldsymbol{\beta}^T X\}\,. \tag{5.15}$$

Furthermore, under Assumptions 3 and 4, τ is identified from the observed data $\mathbf{O}$ as

$$\tau = 1 - \frac{\exp\{\omega + \gamma + \lambda\} - \exp\{\omega\}}{\exp\{\gamma\} - 1}\,. \tag{5.16}$$

Proof. The proof of Proposition 7 can be found in Appendix 5.6.2 □

Remark 4. *The conditions of Proposition 7 lead to an additivity property of the conditional hazard functions that is analogous to Assumption 1. This additivity can be seen in the first equality of* (5.26)*.*

Remark 5. *A weaker version of the randomization / "as-if" randomized condition imposed by Assumption 4 would suffice to identify τ. Namely, that Z, G are independent of all potential outcomes conditional on X.*

Estimation and Inference via the Delta Method

The parameters in (5.15) can be estimated consistently via maximum partial likelihood estimation (Cox, 1975) with the `R` function `coxph` implemented in the `survival` package.

These estimates $(\hat{\omega}, \hat{\gamma}, \hat{\lambda})$ are asymptotically normal with the limiting distribution

$$\sqrt{n}\left(\begin{bmatrix}\hat{\omega}\\ \hat{\gamma}\\ \hat{\lambda}\end{bmatrix} - \begin{bmatrix}\omega\\ \gamma\\ \lambda\end{bmatrix}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(\begin{bmatrix}0\\ 0\\ 0\end{bmatrix}, \mathrm{I}([\omega, \gamma, \lambda^T])^{-1}\right), \tag{5.17}$$

where $\mathrm{I}([\omega, \gamma, \lambda^T])^{-1}$ is the inverse of the partial likelihood-based information matrix. Applying the continuous mapping theorem to (5.16) yields the following consistent MFD estimator of τ,

$$\hat{\tau} = 1 - \frac{\exp\{\hat{\omega} + \hat{\gamma} + \hat{\lambda}\} - \exp\{\hat{\omega}\}}{\exp\{\hat{\gamma}\} - 1}. \tag{5.18}$$

Applying the delta method to (5.17) and (5.18) and noting that $\hat{\mathrm{I}}([\hat{\omega}, \hat{\gamma}, \hat{\lambda}^T])^{-1}/n$, the average sample information, is consistent for $\mathrm{I}([\omega, \gamma, \lambda^T])^{-1}$ gives us an approximate distribution for $\hat{\tau}$ for large enough samples,

$$\hat{\tau} \dot{\sim} \mathcal{N}\left(\tau, \boldsymbol{\partial\hat{\tau}}^T \hat{\mathrm{I}}([\hat{\omega}, \hat{\gamma}, \hat{\lambda}^T])^{-1} \boldsymbol{\partial\hat{\tau}}\right), \tag{5.19}$$

where $\boldsymbol{\partial\hat{\tau}} = [\partial\hat{\tau}/\partial\hat{\omega}, \partial\hat{\tau}/\partial\hat{\gamma}, \partial\hat{\tau}/\partial\hat{\lambda}]^T$. We can use this approximate distribution to conduct inference on τ.

An analogous bounded estimator and confidence interval to those described in Proposition 6 can be constructed using the naive estimator $1 - \exp\{\hat{\omega}\}$. It can easily be shown that this estimator is consistent for a convex combination of τ and η and is thus is a consistent estimator for a lower bound of τ as long as $\eta \leq \tau$.

5.4.2. Causal Identification: Selection Bias and Frailty

Even if these strong modeling assumption assumptions hold, τ defined in §5.4.1 does not have a straightforward causal interpretation. Several authors have discussed in depth the subtleties and common misconceptions of interpreting hazard ratios as a causal quantity (Hernán, 2010; Aalen et al., 2015; Martinussen et al., 2018). We observe one of these

subtleties when examining the definition of τ given in Assumption 6(i),

$$\tau = 1 - \frac{\lim_{\Delta \to 0_+} \mathbb{P}(T^m(1,g) \in [t, t+\Delta)\,|\, T^m(1,g) \geq t,\, X)}{\lim_{\Delta \to 0_+} \mathbb{P}(T^m(0,g) \in [t, t+\Delta)\,|\, T^m(0,g) \geq t,\, X)}\,. \tag{5.20}$$

These authors point out that the hazard functions in the numerator and the denominator of (5.20) condition on different sets of subjects for $t > t_{(1)} = \min_{ij,g,z} T^m_{ij}(g,z)$. The causal contrast is thus some combination of treatment efficacy and a selection effect. Initially, randomized assignment of Z and "as-if" randomization of G ensure that the units in each combination of trial arm and HbAS status are comparable on average. However, conditioning on the study subjects at risk of their first isolated malaria fever at time $t > t_{(1)}$, i.e., $T^m(z,g) \geq t$, has the potential to introduce selection bias. For instance, suppose that baseline susceptibility to malaria is heterogeneous even after conditioning on X. If Z provides protection against developing an isolated malaria fever, we might find that the children with high susceptibility to malaria in the control arm are more likely to develop fevers than similar children in the treatment arm. Consequently, as time passes, the children at risk in the control arm will be less susceptible on average to fever than those in the treatment arm, absent treatment. The comparability of the subjects in each combination of $z, g \in \{0,1\}^2$ ensured by Assumption 4 at the beginning of the study is not guaranteed as the follow-up progresses.

When malaria-attributable fever is rare or when the follow-up time is short, the selection effect may be less consequential (Aalen et al., 2015). If X does not sufficiently capture heterogenous susceptibility to isolated malaria fever, modeling subject-level frailty can further alleviate the selection bias of estimates of τ (Wienke, 2010). frailty can be introduced as a multiplicative random effect,

$$\begin{aligned}
\lambda^m_{zg}(t\,|\,X) &= \lambda(t) W \exp\{\log\alpha + \log(1-\tau)z + \log(1-\nu)g + \boldsymbol{\beta}^T X\}\,, \\
\lambda^{nm}_{zg}(t\,|\,X) &= \lambda(t) W \exp\{\log\phi + \log(1-\eta)z + \boldsymbol{\beta}^T X\}
\end{aligned} \tag{5.21}$$

where W is a random, subject-level frailty shared by both malaria and non-malaria fever hazard functions. Proposition 7 extends immediately to the frailty model implied by (5.21) where, in addition to X, the hazard ratio is conditional on W. Estimation can be carried out via the Expectation-Maximization algorithm (Klein, 1992) or penalized partial maximum likelihood methods (Therneau et al., 2003).

Alternatively, frailty models can be used to conduct a sensitivity analysis of Cox regression-based estimates of τ to selection bias (Stensrud et al., 2017). In general, however, modeling frailty usually requires parametric assumptions about W, for example, that it comes from the family of power variance function distributions (Wienke, 2010).

Although τ itself has a subtle and potentially awkward causal interpretation, simple functions of τ have been shown to have more natural causal interpretations. For example, $1/(2-\tau)$ is the *probabilistic index*, which is defined as the probability that $T^m(1,g)$ for one individual is longer than $T^m(0,g)$ for another individual with comparable baseline covariates (and frailty) (De Neve and Gerds, 2019).

Importantly, all the causal interpretations discussed in this section require the correct identification of the parameter τ established in Proposition 7.

5.5. Discussion

The strategy that we've developed for identification and estimation of malaria vaccine efficacy does not rely on the explicit definition of an inexact, but observable case definition. In short, we propose separating the approach to identifying VE into two distinct steps: first, define a gold-standard case that may be unobservable and then, identify VE using a strategy that doesn't require overt observation of these gold-standard cases. Our gold-standard case definition is one that is 100% specific and sensitive. It is precisely defined using a potential outcome framework. As we've noted, these cases are not distinguishable from observed data alone. Regardless, we demonstrate that with Mendelian factorial design we are able to leverage genetic variation in an analogous fashion to Mendelian randomization studies

to identify VE with respect to this exact, but unobservable, case definition.

In observational studies, *evidence factors* are defined as several approximately independent tests of the same hypothesis, each of which depend on different assumptions about bias from non-random treatment assignment (Rosenbaum, 2011). In the presence of bias, these tests can be thought of as independent pieces of evidence in that the violation of assumptions underlying one test does not imply the other tests are similarly biased. Like an evidence factor, the MFD estimate can provide an additional piece of evidence in vaccine efficacy studies that relies on a different set of assumptions than the methods currently in use. For instance, the naive estimator assumes all fevers with any parasitemia are malaria-attributable, corresponding to a commonly used secondary case definition. Although not independent like true evidence factors, we demonstrated that the naive estimator and the MFD estimator can be combined to construct a *bounded* estimator that outperforms both estimators when used on their own. In particular, the bounded estimator provides significant improvements when the Mendelian gene is only weakly protective, a challenging setting similar to the weak instrument setting in IV studies.

There is evidence that HbAS is moderately protective against uncomplicated malaria and highly protective against severe malaria illness suggesting that MFD may be particularly useful for estimating VE against severe malaria, whose symptoms are not specific and overlap significantly with the symptoms of other severe childhood comorbidities (Bejon et al., 2007).

The performance of the combined estimator across a range of simulation settings with sample sizes that are similar to those found in phase II and III clinical trials is promising. The results suggests that, if feasible, it would be prudent to begin collecting subject data on inherited hemoglobinopathies and other genetic traits that provide protection against clinical malaria, such as the sickle cell trait. Estimating the efficacy of prevention strategies that target transmission directly, such as a malaria transmission-blocking vaccine (Wu et al., 2015) and insecticide-treated bed nets (Ter Kuile et al., 2003), are feasible under the MFD

framework developed in this paper. The assumption of no interference will likely fail but weakening Assumption 3 to allow for partial interference (Hudgens and Halloran, 2008), e.g., interference within but not between villages or sites, is plausible. Cluster randomized trial designs would allow for identification of natural definitions of treatment efficacy that are functions of the fraction of subjects who receive treatment (Athey et al., 2018). The study of erythrocytic vaccines in the MFD framework may be more challenging, as the assumption that protective hemoglobinopathies and these blood-stage vaccines don't interact is tenuous at best. With more than thirty vaccines currently under development, both pre-erythrocytic and those targeting different stages of the disease cycle, the methods described in this paper have the potential to improve the reliability of vaccine efficacy estimates for a number of forthcoming trials (Mahmoudi and Keshavarz, 2017).

Many interesting research directions related to Mendelian factorial design remain. We mention a few in closing. Developing MFD methods using only aggregate site-level data on HbAS prevalence is one such direction. This might allow for MFD-based meta-analyses of past trial results in which hemoglobinopathy data were not collected if, (1) accurate prevalence data could be collected ex post and (2) the prevalence varied sufficiently between studies. There is evidence that there are a number of other genetic traits that confer protection against malaria (Ndila et al., 2018). When there are several potential valid Mendelian factors, we may again find inspiration from MR and other IV methods. It would be beneficial to develop falsification tests of the validity of potential Mendelian factors akin to tests for over-identifying restrictions like the Sargan-Hansen test in IV regression (Hansen, 1982). As it has been demonstrated in MR, using many ostensible Mendelian factors where some are invalid may have the potential to improve the robustness of MFD analysis (Kang et al., 2016).

5.6. Appendix

5.6.1. Proof of Proposition 4

In this Appendix we give a proof of the general identification result in Proposition 4.

Proof of Proposition 4. By consistency and Assumption 3 we have that

$$\mathbb{E}_P[f(Y)|\, X, Z = z, G = g] = \mathbb{E}_P[f(Y(z,g))|\, X, Z = z, G = g] \tag{5.22}$$

for all z, g. Assumption 4 ensures that $P_{X|Z,G} = P_X$ and $P_{Y(z,g)|Z,G} = P_{Y(z,g)}$. Hence,

$$\begin{aligned}
\mathbb{E}_X[\mathbb{E}_P[f\{Y(z,g)\}|\, X, Z = z, G = g]] &= \mathbb{E}_{X|Z=z,G=g}[\mathbb{E}_P[f\{Y(z,g)\}|\, X, Z = z, G = g]] \\
&= \mathbb{E}_P[f\{Y(z,g)\}|\, Z = z, G = g] \\
&= \mathbb{E}_P[f\{Y(z,g)\}] \\
&= \mu_{zg}(P)\,.
\end{aligned} \tag{5.23}$$

Applying Assumption 1, the right hand side of (5.2) becomes

$$1 - \frac{(\mu^m_{11}(P) - \mu^m_{10}(P)) + (\mu^{nm}_{11}(P) - \mu^{nm}_{10}(P))}{(\mu^m_{01}(P) - \mu^m_{00}(P)) + (\mu^{nm}_{01}(P) - \mu^{nm}_{00}(P))}\,. \tag{5.24}$$

Because G is a valid Mendelian factor (Assumption 5(ii)) we have that $\mu^{nm}_{z1}(P) - \mu^{nm}_{z0}(P) = 0$ for $z = 0, 1$, further simplifying (5.24) to

$$1 - \frac{(\mu^m_{11}(P) - \mu^m_{10}(P))}{(\mu^m_{01}(P) - \mu^m_{00}(P))}\,.$$

Finally, we have that

$$\begin{aligned} 1 - \frac{(\mu_{11}^m(P) - \mu_{10}^m(P))}{(\mu_{01}^m(P) - \mu_{00}^m(P))} &= 1 - \frac{-\nu\mu_{10}^m(P)}{-\nu\mu_{00}^m(P)} && \text{by Assumption 2 and Definition 4(ii)} \\ &= 1 - \frac{\mu_{10}^m(P)}{\mu_{00}^m(P)} \\ &= \tau && \text{by Definition 4(i)}\,. \end{aligned}$$

Assumption 5(i) ensures that the right hand side of the first equality is well-defined. □

5.6.2. Proof of Proposition 6

We give a short proof of the asymptotic unbiasedness of $\hat{\tau}_{bnd}$ and the asymptotical validity of $\text{CI}_{bnd,\alpha}$ as a $1-\alpha$ confidence interval for τ.

Proof of Proposition 6. We give a proof for when the sample sizes are balanced across sites. When this is the case, we skip step 2 of Algorithm 1 and use $\hat{\mu}_0$ to construct $\hat{\tau}_0$. Treating G as an element of X, the consistency of $\hat{\tau}_0$ for $s\tau+(1-s)\eta$ and its asymptotic linearity follows almost immediately from Theorem 1 of Rosenblum and van der Laan (2010). When $\eta < \tau$, the theorem implies that $\hat{\tau}_0$ is consistent for some $\tau' < \tau$. Because $\tau \leq 1$ by definition, the continuous mapping theorem implies that $\hat{\tau}_{bnd}$ is consistent for $\min[1, \max\{\tau', \tau\}] = \tau$. The lower bound of $\text{CI}_{bnd,\alpha}$ is constructed by taking the intersection of a lower one-sided confidence interval for τ with coverage $1 - \alpha/2 + \alpha_0$ and a lower one-sided confidence interval for τ' with coverage $1 - \alpha_0$. Because $\tau' < \tau$, this second confidence interval is also asymptotically valid for τ. If α_0 is chosen a priori, then the intersection is an asymptotically valid $1-\alpha/2$ lower one-sided interval for τ (Neuwald and Green, 1994). The upper confidence bound is constructed by taking the intersection of a $1 - \alpha/2$ upper one-sided confidence interval for τ and $(-\infty, 1]$. Since $\tau \leq 1$ the resulting interval is an asymptotically valid $1 - \alpha/2$ upper confidence-interval for τ. The intersection of the resultant upper and lower $1 - \alpha/2$ one-sided confidence intervals yields an asymptotically valid two-sided confidence interval with coverage $1 - \alpha$. □

5.6.3. Proof of Proposition 7

In this Appendix we provide a proof of the parameter identification result for the hazard ratio under the proportional hazards assumption.

Proof of Proposition 7. Let $\Lambda_{zg}(t \mid X) = \int_{-\infty}^{t} \lambda_{zg}(t \mid X)\, dt$ be the cumulative hazard function for Y. Define $\Lambda_{zg}^{k}(t \mid X)$, $k = m, nm$ similarly. The survival function $S_{zg}(t \mid X = \mathbb{P}(T(z,g) > t \mid X)$ can be expressed as $S_{zg}(t \mid X) = \exp\{-\Lambda_{zg}(t \mid X)\}$ and the probability density of Y as $f_{zg}(t \mid X) = \lambda_{zg}(t \mid X) S_{zg}(t \mid X)$. We can express $f_{zg}(t \mid X)$ as

$$\begin{aligned}
f_{zg}(t \mid X) &= \frac{\partial}{\partial t} F_{zg}(t \mid X) \\
&= \frac{\partial}{\partial t} \left\{1 - S_{zg}^{nm}(t \mid X) S_{zg}^{m}(t \mid X)\right\} \quad \text{by Assumption 7} \\
&= f_{zg}^{m}(t \mid X) S_{zg}^{nm}(t \mid X) + f_{zg}^{nm}(t \mid X) S_{zg}^{m}(t \mid X) \\
&= \{\lambda_{zg}^{m}(t \mid X) + \lambda_{zg}^{nm}(t \mid X)\} S_{zg}^{nm}(t \mid X) S_{zg}^{m}(t \mid X) \\
&= \{\lambda_{zg}^{m}(t \mid X) + \lambda_{zg}^{nm}(t \mid X)\} \times \exp\left\{ - \int_{-\infty}^{t} \lambda_{zg}^{m}(t \mid X) + \lambda_{zg}^{nm}(t \mid X)\, dt \right\}. \quad (5.25)
\end{aligned}$$

From (5.25) and Assumptions 6 and 8, we have that the hazard function for Y is

$$\begin{aligned}
\lambda_{zg}(t \mid X) &= \lambda_{zg}^{m}(t \mid X) + \lambda_{zg}^{nm}(t \mid X) \\
&= \lambda(t) \exp\{\boldsymbol{\beta}^T X\} \left(\kappa(1-\tau)^z (1-\nu)^g + (1-\eta)^z\right) \\
&= \lambda(t) \exp\{\alpha + \omega z + \gamma g + \lambda z \times g + \boldsymbol{\beta}^T X\}, \quad (5.26)
\end{aligned}$$

Proving the first part of the proposition. To prove the second part of the proposition we first observe that the last equality above in (5.26) is simply a reparameterization – when z and g are binary, $\kappa(1-\tau)^z(1-\nu)^g + \phi(1-\eta)^z$ and $\exp\{\alpha + \omega z + \gamma g + \lambda z \times g\}$ can each

take four distinct values. The reparameterization yields a system of four equations,

$$\exp\{\alpha\} = \kappa + \phi \tag{5.27}$$

$$\exp\{\alpha + \omega\} = \kappa(1-\tau) + \phi(1-\eta) \tag{5.28}$$

$$\exp\{\alpha + \gamma\} = \kappa(1-\nu) \tag{5.29}$$

$$\exp\{\alpha + \omega + \gamma + \lambda\} = \kappa(1-\tau)(1-\nu) + \phi(1-\eta)\,. \tag{5.30}$$

Subtracting (5.28) from (5.30) and (5.27) from (5.29), then dividing the former by the latter yields

$$1 - \tau = \frac{\exp\{\omega + \gamma + \lambda\} - \exp\{\omega\}}{\exp\{\gamma\} - 1}\,. \tag{5.31}$$

Under Assumptions 3 and 4 we have that $\lambda_{zg}(t \,|\, X)$ can be identified by the observed data $\mathbf{O}$:

$$\begin{aligned}\lambda_{zg}(t \,|\, X) &= \lim_{\Delta \to 0_+} \mathbb{P}(T(z,g) \in [t, t+\Delta) | \, T(z,g) \geq t,\, X) \,/\Delta \\ &= \lim_{\Delta \to 0_+} \mathbb{P}(T \in [t, t+\Delta) | \, T \geq t,\, X,\, Z = z,\, G = g) \,/\Delta\,. \end{aligned} \tag{5.32}$$

Applying (5.26), we have that $\alpha, \omega, \gamma, \lambda$ are identifiable and thus τ can be identified by the observed data using (5.31). □

5.6.4. Proof (Sketch) of Proposition 5

In this section we sketch a proof of Proposition 5 by showing that $\mu_{zg}(P_n)$ is consistent for $\mu_{zg}(P)$ and asymptotically linear, that is,

$$\mu_{zg}(P_n) - \mu_{zg}(P) = (\mathbb{P}_n - \mathbb{P})\varphi^*_{zg}(P_\infty) + o_P\left(1/\sqrt{N}\right)\,. \tag{5.33}$$

for all z, g. Theorem 1 of Rosenblum and van der Laan (2010) checks a number of technical conditions to verify that Theorem 1 of van der Laan and Rubin (2006) can be applied. A particularly important condition is that the $\mu_{zg}(P)$ are linear. This does not hold in the

setting when the prevalence of G in each site is not known a priori. Hence, a straightforward application of the theorem is not appropriate. However, if certain weaker conditions hold, then the important parts of Theorem 1 of van der Laan and Rubin (2006) still apply. We first show that we can write

$$\mu_{zg}(P_n) - \mu_{zg}(P) = (\mathbb{P}_n - \mathbb{P})\varphi^*_{zg}(P_n)\,. \tag{5.34}$$

If $\varphi^*_{zg}(P_n)$ is Donsker then by the second part of Theorem 1 of van der Laan and Rubin (2006), we have that $\mu_{zg}(P_n)$ is $\sqrt{n}$-consistent for $\mu_{zg}(P)$. Then, if $\mathbb{P}\{\varphi^*_{zg}P_n - \varphi^*_{zg}(P_\infty)\}^2 \to 0$ in probability as $n \to \infty$ the third part of Theorem 1 of van der Laan and Rubin (2006) implies that $\mu_{zg}(P_n)$ is asymptotically linear with form (5.33). Consistency of $\mu_{zg}(P_n)$ and Assumptions 1 - 5 imply that the plugin estimator $\hat{\tau}$ is consistent for τ. Finally, we apply the delta method to derive the asymptotic variance σ^2 of $\sqrt{n}(\hat{\tau} - \tau)$. It follows from the last part of Theorem 1 of van der Laan and Rubin (2006) that σ^2 achieves the semiparametric efficiency bound if the working model for the conditional expectation of $f(Y)$ is correctly specified. In what follows, when taking expectations of functionals of P_n we treat these functionals as fixed.

Verifying (5.34)*:*

Given the choice of the terms included in the linear part of the GLM estimate $\hat{\mu}_1$ in Algorithm 1, the score equations that $\hat{\mu}_1$ solve imply that

$$\mathbb{P}_n\left\{\frac{\mathbb{1}(Z=z)\mathbb{1}(G=g)}{p(Z=z)p_{j,n}(G=g)}\left(f(Y) - \hat{\mu}_1(z,g,X)\right)\right\} = 0 \;\text{ for all }\; z,g \in \{0,1\}^2\,. \tag{5.35}$$

The choice of the weighting $\mathtt{w} = n/I_j$ when estimating $\hat{\mu}_1$ is required since $\mathbb{P}_n$ does not equally weight observations unless the size of each site is the same. Also note, that by definition, $\mathbb{P}_n\hat{\mu}_1(z,g,X) = \mu_{zg}(P_n)$. Taken together, we have that $\mathbb{P}_n\varphi_{zg}(P_n) = 0$ for all

$z, g \in \{0,1\}^2$. By iterated expectations, we have that

$$\mathbb{P}\varphi^*_{zg}(P_n) = \mathbb{P}\varphi_{zg}(P_n) - \mathbb{P}\left\{\mathbb{E}_P[\varphi_{zg}(P_n) \,|\, G]\right\} = \mathbb{P}\varphi_{zg}(P_n) - \mathbb{P}\varphi_{zg}(P_n) = 0\,. \tag{5.36}$$

All that remains to be shown is that $-\mathbb{P}_n\left\{\mathbb{E}_P[\varphi_{zg}(P_n) \,|\, G]\right\} = \mu_{zg}(P_n) - \mu_{zg}(P)$. We start by evaluating $\mathbb{E}_P[\varphi_{zg}(P_n) \,|\, G]$:

$$\begin{aligned}
\mathbb{E}_P[\varphi_{zg}(P_n) \,|\, G] &= \mathbb{E}_P\left[\frac{\mathbb{1}(Z=z)\mathbb{1}(G=g)\left\{f(Y) - \hat{\mu}_1(z,g,X)\right\}}{p_{j,n}(G=g)p(Z=z)} \,\Big|\, G\right] \\
&\quad + \mathbb{E}_P[\hat{\mu}_1(z,g,X) - \mu_{zg}(P_n) \,|\, G] \\
&= \frac{\mathbb{1}(G=g)}{p_{j,n}(G=g)}\left(\mu_{zg}(P) - \mathbb{E}_P[\hat{\mu}_1(z,g,X) \,|\, G]\right) \\
&\quad + \mathbb{E}_P[\hat{\mu}_1(z,g,X) \,|\, G] - \mu_{zg}(P_n)\,.
\end{aligned} \tag{5.37}$$

The second equality follows from an application of iterated expectations, Assumption 4, and the fact that $\mu_{zg}(P_n)$ and $\hat{\mu}_1$ are treated as fixed when taking expectations. Taking the empirical expectation $\mathbb{P}_n$ we get

$$\begin{aligned}
\mathbb{P}_n\left\{\mathbb{E}_P[\varphi_{zg}(P_n) \,|\, G]\right\} &= \mathbb{P}_n\left\{\frac{\mathbb{1}(G=g)}{p_{j,n}(G=g)}\left(\mu_{zg}(P) - \mathbb{E}_P[\hat{\mu}_1(z,g,X) \,|\, G]\right)\right\} \\
&\quad + \mathbb{P}_n\left\{\mathbb{E}_P[\hat{\mu}_1(z,g,X) \,|\, G] - \mu_{zg}(P_n)\right\} \\
&= \mu_{zg}(P) - \mathbb{E}_P[\hat{\mu}_1(z,g,X) \,|\, G=g] \\
&\quad + \mathbb{P}_n\left\{\mathbb{E}_P[\hat{\mu}_1(z,g,X) \,|\, G]\right\} - \mu_{zg}(P_n) \\
&= \mu_{zg}(P) - \mathbb{E}_P[\hat{\mu}_1(z,g,X)] + \mathbb{E}_P[\hat{\mu}_1(z,g,X)] - \mu_{zg}(P_n) \\
&= -(\mu_{zg}(P_n) - \mu_{zg}(P))\,.
\end{aligned} \tag{5.38}$$

The second to last equality comes from the fact that $X \perp\!\!\!\perp G$ by Assumption 4. This is our desired result and (5.34) now follows.

Checking that $\varphi^*_{zg}(P_n)$ is Donsker:

This is analogous to condition (iv) in the proof of Theorem 1 in Rosenblum and van der Laan (2010). We make the same assumptions of boundedness on $\boldsymbol{\beta}$ and on the terms in the linear part of the generalized linear models in steps 1 and 2 of Algorithm 1. Under the additional assumption that $0 < \delta \leq p_j(G = g)$ for all $g = 0, 1$ and $j = 1, \ldots, J$, a nearly identical verification that $\varphi^*_{zg}(P_n)$ is Donsker follows. Hence, $\mu_{zg}(P_n)$ are $\sqrt{n}$-consistent for all $z, g \in \{0, 1\}1^2$.

Verifying that $\mathbb{P}\{\varphi^*_{zg}(P_n) - \varphi^*_{zg}(P_\infty)\}^2 \xrightarrow{P} 0$:

The steps to verify that $\mathbb{P}\{\varphi^*_{zg}(P_n) - \varphi^*_{zg}(P_\infty)\}^2 \to 0$ in probability are very similar to the verification of condition (v) in Rosenblum and van der Laan (2010). There are a few extra steps required to deal with the fact that we are estimating $p_j(G = g)$ with $p_{j,n}(G = g)$. We first note that we can write

$$\begin{aligned}
\mathbb{P}\{\varphi^*_{zg}P_n - \varphi^*_{zg}(P_\infty)\}^2 &= \mathbb{P}[\{\varphi_{zg}(P_n) - \varphi_{zg}(P_\infty)\} - \mathbb{E}_P[\varphi_{zg}(P_\infty) - \varphi_{zg}(P_n) \,|\, G]]^2 \\
&\leq 2\mathbb{P}\{\varphi_{zg}(P_n) - \varphi_{zg}(P_\infty)\}^2 + 2\mathbb{P}\{\mathbb{E}_P[\varphi_{zg}(P_\infty) - \varphi_{zg}(P_n) \,|\, G]\}^2 \\
&\leq 4\mathbb{P}\{\varphi_{zg}(P_n) - \varphi_{zg}(P_\infty)\}^2 . \qquad (5.39)
\end{aligned}$$

The last line comes from applying Jenson's inequality to the inner expectation of the second term followed by an application of iterated expectations. We need to show that $\mathbb{P}\{\varphi_{zg}(P_n) - \varphi_{zg}(P_\infty)\}^2$ converges to 0 in probability. In what follows we distinguish the working model using the fitted parameters $\boldsymbol{\beta}_n$ and the model using the unique maximizer of the expected log-likelihood $\boldsymbol{\beta}$ as $\hat{\mu}_{1,n}$ and $\hat{\mu}_{1,\infty}$, respectively. For brevity, we also let $\mathbb{1}_{zg} = \mathbb{1}(Z =$

$z)\mathbb{1}(G=g)$, $p(z)=p(Z=z)$, $p_j(g)=p_j(G=g)$, and $p_{j,n}(g)=p_{j,n}(G=g)$. We can write

$$
\begin{aligned}
&\mathbb{P}\{\varphi_{zg}(P_n)-\varphi_{zg}(P_\infty)\}^2 \\
&\quad=\mathbb{P}\left\{\frac{\mathbb{1}_{zg}}{p(z)p_{j,n}(g)}\{f(Y)-\hat{\mu}_{1,n}(z,g,X)\}+\hat{\mu}_{1,n}(z,g,X)-\mu_{zg}(P_n)\right. \\
&\qquad\left.-\frac{\mathbb{1}_{zg}}{p(z)p_j(g)}\{f(Y)-\hat{\mu}_{1,\infty}(z,g,X)\}-\hat{\mu}_{1,\infty}(z,g,X)+\mu_{zg}(P)\right\}^2 \\
&\quad\leq\frac{4}{p(z)^2}\mathbb{P}\left\{f(Y)^2\right\}\left(\frac{1}{p_{j,n}(g)}-\frac{1}{p_j(g)}\right)^2+\frac{4}{p(z)^2}\mathbb{P}\left\{\frac{\hat{\mu}_{1,\infty}(z,g,X)}{p_j(g)}-\frac{\hat{\mu}_{1,n}(z,g,X)}{p_{j,n}(g)}\right\}^2 \\
&\qquad+4\mathbb{P}\left\{\hat{\mu}_{1,n}(z,g,X)-\hat{\mu}_{1,\infty}(z,g,X)\right\}^2+4\mathbb{P}\left\{\mu_{zg}(P)-\mu_{zg}(P_n)\right\}^2 \\
&\quad\leq C_1\left(\frac{1}{p_{j,n}(g)}-\frac{1}{p_j(g)}\right)^2+\frac{4}{p(z)^2}\mathbb{P}\left\{\frac{\hat{\mu}_{1,\infty}(z,g,X)}{p_j(g)}-\frac{\hat{\mu}_{1,n}(z,g,X)}{p_{j,n}(g)}\right\}^2 \\
&\qquad+C_2\|\boldsymbol{\beta}_0-\boldsymbol{\beta}\|^2+4\{\mu_{zg}(P)-\mu_{zg}(P_n)\}^2\,. \qquad (5.40)
\end{aligned}
$$

The first equality follows immediately from definitions. The first *inequality* follows from rearranging terms and noting that $(x_1+\cdots+x_k)^2\leq k(x_1^2+\cdots+x_k^2)$ by Jensen's inequality. The first term after the second inequality comes from the boundedness of Y. Since $p_{j,n}(g)\xrightarrow{P}p_j(g)$ and we have assumed that $p_j(g)$ are bounded away from zero, the continuous mapping theorem implies that this term converges to 0 in probability. The third term comes from the fact that our working model has uniformly bounded first derivatives. This is due to the boundedness assumptions on X and the terms of the linear part of our working model. The conditions given in Proposition 5, are sufficient for $\boldsymbol{\beta}_n$ to converge to $\boldsymbol{\beta}$ in probability (Rosenblum and van der Laan (2009), Appendix D). Consequently, the third term also converges to 0 in probability. The fourth term disappears in probability since we proved earlier that $\mu_{zg}(P_n)$ is consistent for $\mu_{zg}(P)$. All that is left to handle is the second term.

A little bit of rearranging gives us

$$\begin{aligned}
&\frac{4}{p(z)^2}\mathbb{P}\left\{\frac{\hat{\mu}_{1,\infty}(z,g,X)}{p_j(g)} - \frac{\hat{\mu}_{1,n}(z,g,X)}{p_{j,n}(g)}\right\}^2 \\
&\quad= \frac{4}{p(z)^2}\mathbb{P}\left\{\frac{\hat{\mu}_{1,\infty}(z,g,X) - \hat{\mu}_{1,n}(z,g,X)}{p_j(g)} + \hat{\mu}_{1,n}(z,g,X)\left(\frac{1}{p_{j,n}(g)} - \frac{1}{p_j(g)}\right)\right\}^2 \\
&\quad\leq \frac{8}{\{p(z)p_j(g)\}^2}\mathbb{P}\left\{\hat{\mu}_{1,\infty}(z,g,X) - \hat{\mu}_{1,n}(z,g,X)\right\}^2 \\
&\qquad+ \frac{8}{p(z)^2}\mathbb{P}\left\{\hat{\mu}_{1,n}(z,g,X)^2\right\}\left(\frac{1}{p_{j,n}(g)} - \frac{1}{p_j(g)}\right)^2 \\
&\quad= C_3\|\boldsymbol{\beta}_0 - \boldsymbol{\beta}\|^2 + C_4\left(\frac{1}{p_{j,n}(g)} - \frac{1}{p_j(g)}\right)^2 . \qquad (5.41)
\end{aligned}$$

The first inequality follows again from Jenson's inequality. The first term in the last line follows from the same arguments made earlier and the second term follows from the fact that $\mathbb{P}\left\{\hat{\mu}_{1,n}(z,g,X)^2\right\}$ is bounded. As we've already demonstrated, these two terms vanish in probability. Combining (5.39), (5.40), and (5.41) gives us that $\mathbb{P}\{\varphi^*_{zg}(P_n) - \varphi^*_{zg}(P_\infty)\}^2 \xrightarrow{P} 0$. (5.35) follows immediately and an application of Proposition 4 implies that $\hat{\tau}$ is also asymptotically linear. All that remains is to compute its asymptotic variance.

Asymptotic Variance:

To compute the asymptotic variance of $\hat{\tau}$ we begin with a taylor expansion around τ:

$$\begin{aligned}
\hat{\tau} &= 1 - \frac{\mu_1(P_n)}{\mu_0(P_n)} \\
&= 1 - \frac{\mu_1(P)}{\mu_0(P)} - \frac{1}{\mu_0(P)}\{\mu_1(P_n) - \mu_1(P)\} \\
&\qquad+ \frac{\mu_1(P)}{\mu_0(P)^2}\{\mu_0(P_n) - \mu_0(P)\} + o_p\left(1/\sqrt{n}\right) \\
&= \tau - \frac{1}{\mu_0(P)}(\mathbb{P}_n - \mathbb{P})\varphi^*_1(P_\infty)(O) + \frac{\mu_1(P)}{\mu_0(P)^2}(\mathbb{P}_n - \mathbb{P})\varphi^*_0(P_\infty)(O) + o_p(1/\sqrt{n})\,. \qquad (5.42)
\end{aligned}$$

The $o_p(1/\sqrt{n})$ comes from the second order remainder term of the expansion. Using the expansion in (5.42) along with the fact that $\varphi^*_z(P_\infty)(O) - \mathbb{P}\varphi^*_z(P_\infty)(O)$ are mean zero i.i.d.

random variables we can write the scaled asymptotic variance of $\hat{\tau}$, $n \cdot \text{var}(\hat{\tau})$, as

$$\mathbb{E}_P \left[\frac{\mu_1(P)}{\mu_0(P)^2} (\varphi_0^*\{P_\infty\})(O) - \mathbb{P}\varphi_0^*(P_\infty)(O)\} - \frac{1}{\mu_0(P)} \{\varphi_1^*(P_\infty)(O) - \mathbb{P}\varphi_1^*(P_\infty)(O)\} \right]^2 . \quad (5.43)$$

5.6.5. Additional Simulation Study Details for TMLE-based Estimators

Below are the simulation settings for the individual-specific vaccine and protective efficacies (τ_i and ν_i), the baseline covariates and how they are transformed when they enter the condition mean of the distribution of fever counts (X_i, $\tilde{X}_i^m$, and $\tilde{X}_i^{nm}$), and unobserved heterogeneity in the distribution of fever counts between individuals (ϵ_i^m and ϵ_i^{nm}).

Generative Distributions:

$$\begin{aligned}
(1 - \nu_i) &\sim (1 - \nu) \exp\{-0.05^2/2\} \cdot \texttt{Lognormal}(0, 0.05^2) \\
(1 - \tau_i) &\sim (1 - \tau) \exp\{-0.05^2/2\} \cdot \texttt{Lognormal}(0, 0.05^2) , \\
\tilde{X}_i^m &\sim \exp\{0.05 X_i - 0.05^2/2\} , \\
\tilde{X}_i^{nm} &\sim \exp\{0.075 X_i - 0.075^2/2\} , \\
X_i &\sim \texttt{Normal}(0, 1) , \\
\epsilon_i^m &\sim \exp\{-0.05^2/2\} \cdot \texttt{Lognormal}(0, 0.05^2) , \\
\epsilon_i^{nm} &\sim \exp\{-0.05^2/2\} \cdot \texttt{Lognormal}(0, 0.05^2) .
\end{aligned}$$

We assume that the spillover efficacy η is equal to zero. Based on empirical evidence and the negative dependence between malaria-attributable fevers and non-malaria fevers as defined in (5.1), we simulate the number of malaria-attributable and non-malaria fevers from negatively dependent negative binomial distributions (Olotu et al., 2013). We use a overdispersion parameter of $r = 10$

Count of malaria-attributable Fevers per child-year:

On the margin, the count of malaria-attributable fevers per child-year follows a negative binomial distribution with mean $\mu_{m,i}$ and variance $\sigma^2_{m,i}$,

$$\mathbf{1}^T Y_i^m(z,g) \sim \texttt{NB}(\mu_{m,i}, \sigma^2_{m,i}) \tag{5.44}$$

where $\mu_{m,i} = \kappa \cdot (1-\nu_i)^g \cdot (1-\tau_i)^z \cdot \tilde{X}_i^m \cdot \epsilon_i^m$ and $\sigma^2_{m,i} = \mu^2_{m,i}/r + \mu_{m,i}$.

Count of non-malaria fevers per child-year:

On the margin, the count of non-malaria fevers per child-year follows a negative binomial distribution with mean $\mu_{nm,i}$ and variance $\sigma^2_{nm,i}$,

$$\mathbf{1}^T Y_i^{nm}(z,g) \sim \texttt{NB}(\mu_{nm,i}, \sigma^2_{nm,i}) \tag{5.45}$$

where $\mu_{nm,i} = \phi \cdot \tilde{X}_i^{nm} \cdot \epsilon_i^{nm}$ and $\sigma^2_{nm,i} = \mu^2_{nm,i}/r + \mu_{nm,i}$.

Count of fevers of any-cause per child-year:

The joint distribution of the counts of non-malaria and malaria-attributable fevers per child-year are simulated using a Gaussian copula with a negative dependence parameter $\rho = -0.1$ (Genest and Neslehová, 2007).

Specificity and Mendelian Gene Prevalence Settings:

We calibrated κ and ϕ to achieve different levels of case specificity ($s = 0.5$ and 0.8). For example, for a case specificity of 0.8 we set κ and ϕ so that the expected number of malaria-attributable fevers is 80% of the expected number of total fevers of any-cause. The prevalence of the Mendelian gene was set to 20% based on existing estimates of HbAS

prevalence in sub-Saharan Africa (Ter Kuile et al., 2003; Elguero et al., 2015).

Initial and Updated Working Model Estimates:

For both the MFD and naive estimators, we used the `R` function `glm.fit` from the package `stats` to fit the initial estimator $\hat{\mu}_0$ in step 1 of Algorithm 1 using Poisson regression. The regression included main terms for Z, G, and X as well as all interactions.

Bibliography

Aalen, O. O., Cook, R. J., and Røysland, K. (2015). Does cox analysis of a randomized survival study yield a causal treatment effect? *Lifetime data analysis*, 21(4):579–593.

Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*, 105(490):493–505.

Ashenfelter, O. and Rouse, C. (1998). Income, schooling, and ability: Evidence from a new sample of identical twins. *The Quarterly Journal of Economics*, 113(1):253–284.

Athey, S., Eckles, D., and Imbens, G. W. (2018). Exact p-values for network interference. *Journal of the American Statistical Association*, 113(521):230–240.

Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.

Becker, G. S. (2009). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.

Bejon, P., Berkley, J. A., Mwangi, T., Ogada, E., Mwangi, I., Maitland, K., Williams, T., Scott, J. A. G., English, M., Lowe, B. S., et al. (2007). Defining childhood severe falciparum malaria for intervention studies. *PLoS medicine*, 4(8):e251.

Bejon, P., White, M. T., Olotu, A., Bojang, K., Lusingu, J. P., Salim, N., Otsyula, N. N., Agnandji, S. T., Asante, K. P., Owusu-Agyei, S., et al. (2013). Efficacy of rts, s malaria vaccines: individual-participant pooled analysis of phase 2 data. *The Lancet infectious diseases*, 13(4):319–327.

Berger, R. L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300.

Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016.

Bernal, J. L., Cummins, S., and Gasparrini, A. (2017). Interrupted time series regression for the evaluation of public health interventions: a tutorial. *International journal of epidemiology*, 46(1):348–355.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275.

Bhatia, R. and Davis, C. (2000). A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357.

Blitstein, J. L., Hannan, P. J., Murray, D. M., and Shadish, W. R. (2005). Increasing the degrees of freedom in existing group randomized trials: The df* approach. *Evaluation Review*, 29(3):241–267.

Brillinger, D. R. (1986). A biometrics invited paper with discussion: the natural variability of vital rates and associated statistics. *Biometrics*, 42:693–734.

Burgess, S., Small, D. S., and Thompson, S. G. (2017). A review of instrumental variable estimators for mendelian randomization. *Statistical methods in medical research*, 26(5):2333–2355.

Burstein, A. and Vogel, J. (2017). International trade, technology, and the skill premium. *Journal of Political Economy*, 125(5):1356–1412.

Campbell, D. T. (1969). Prospective: Artifact and control. In Rosenthal, R. and Rosnow, R. L., editors, *Artifacts in Behavioral Research: Robert Rosenthal and Ralph L. Rosnow's Classic Books*, pages 351–382. Academic Press, New York, NY.

Campbell, D. T. and Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In Gage, N., editor, *Handbook of research on teaching*, pages 171–246. Rand McNally, Chicago, IL.

Card, D. (1999). The causal effect of education on earnings. volume 3 of *Handbook of Labor Economics*, pages 1801 – 1863. Elsevier.

Card, D. and Krueger, A. B. (1992). Does school quality matter? Returns to education and the characteristics of public schools in the United States. *Journal of Political Economy*, 100(1):1–40.

Conley, T. G. and Taber, C. R. (2011). Inference with "difference in differences" with a small number of policy changes. *The Review of Economics and Statistics*, 93(1):113–125.

Cook, T. D. and Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Rand McNally, Chicago, IL.

Cook, T. D., Campbell, D. T., and Shadish, W. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin, Boston, MA.

Cornfield, J., Haenszel, W., Hammond, E. C., Lilienfeld, A. M., Shimkin, M. B., and Wynder, E. L. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1):173–203.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Crifasi, C. K., Merrill-Francis, M., McCourt, A., Vernick, J. S., Wintemute, G. J., and Webster, D. W. (2018). Association between firearm laws and homicide in urban counties. *Journal of urban health*, 95(3):383–390.

De Neve, J. and Gerds, T. A. (2019). On the interpretation of the hazard ratio in cox regression. *Biometrical Journal*.

Dix-Carneiro, R. and Kovak, B. K. (2017). Trade liberalization and regional dynamics. *American Economic Review*, 107(10):2908–46.

Donald, S. G. and Lang, K. (2007). Inference with difference-in-differences and other panel data. *The Review of Economics and Statistics*, 89(2):221–233.

Donovan, S. J. and Susser, E. (2011). Commentary: Advent of sibling designs. *International Journal of Epidemiology*, 40(2):345.

Egleston, B. L., Scharfstein, D. O., and MacKenzie, E. (2009). On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics*, 65(2):497–504.

Elguero, E., Délicat-Loembet, L. M., Rougeron, V., Arnathau, C., Roche, B., Becquart, P., Gonzalez, J.-P., Nkoghe, D., Sica, L., Leroy, E. M., et al. (2015). Malaria continues to select for sickle cell trait in central africa. *Proceedings of the National Academy of Sciences*, 112(22):7051–7054.

Ferreira, A., Marguti, I., Bechmann, I., Jeney, V., Chora, Â., Palha, N. R., Rebelo, S., Henri, A., Beuzard, Y., and Soares, M. P. (2011). Sickle hemoglobin confers tolerance to plasmodium infection. *Cell*, 145(3):398–409.

Fine, J. P. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509.

Fisher, R. A. (1935). *The Design of Experiments*. Oliver & Boyd.

Fogarty, C. B. and Hasegawa, R. B. (2019). Extended sensitivity analysis for heterogeneous unmeasured confounding with an application to sibling studies of returns to education. to appear.

Fogarty, C. B. and Small, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *Journal of the American Statistical Association*, 111(516):1820–1830.

Frisell, T., Öberg, S., Kuja-Halkola, R., and Sjölander, A. (2012). Sibling comparison designs: Bias from non-shared confounders and measurement error. *Epidemiology*, 23(5):713–720.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*, 85(4):pp. 907–920.

Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (2000). Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(3):545–555.

Genest, C. and Neslehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, 37(2):475–515.

Gleser, L. J. (1975). On the distribution of the number of successes in independent trials. *Ann. Probab.*, 3(1):182–188.

Gong, L., Parikh, S., Rosenthal, P. J., and Greenhouse, B. (2013). Biochemical and immunological mechanisms by which sickle cell trait protects against malaria. *Malaria Journal*, 12(1):317.

Gravenor, M. and Kwiatkowski, D. (1998). An analysis of the temperature effects of fever on the intra-host population dynamics of plasmodium falciparum. *Parasitology*, 117(2):97–105.

Griliches, Z. (1970). Notes on the role of education in production functions and growth accounting. In *Education, Income, and Human Capital*, NBER Chapters, pages 71–127. National Bureau of Economic Research, Inc.

Griliches, Z. (1979). Sibling models and data in economics: Beginnings of a survey. *Journal of Political Economy*, 87(5):S37–S64.

Hájek, J., Šidák, Z., and Sen, P. K. (1999). *Theory of Rank Tests*. Academic Press, San Diego.

Hammond, E. C. (1964). Smoking in relation to mortality and morbidity. findings in first thirty-four months of follow-up in a prospective study started in 1959. *Journal of the National Cancer Institute*, 32(5):1161–1188.

Hansen, B. B. and Klopfer, S. O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627.

Hansen, C. B. (2007). Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of Econometrics*, 140(2):670–694.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054.

Hasegawa, R. and Small, D. (2017). Sensitivity analysis for matched pair analysis of binary data: From worst case to average case analysis. *Biometrics*, 73(4):1424–1432.

Hasegawa, R. B., Webster, D. W., and Small, D. S. (2019). Evaluating missouri's handgun purchaser law. *Epidemiology*, 30(3):371–379.

Hauser, R. M., Sheridan, J. T., and Warren, J. R. (1999). Socioeconomic achievements of siblings in the life course. *Research on Aging*, 21(2):338–378.

Herd, P., Carr, D., and Roan, C. (2014). Cohort profile: Wisconsin longitudinal study (WLS). *International Journal of Epidemiology*, 43(1):34–41.

Hernán, M. A. (2010). The hazards of hazard ratios. *Epidemiology*, 21(1):13–15.

Hoeffding, W. (1956). On the distribution of the number of successes in independent trials. *Ann. Math. Statist.*, 27(3):713–721.

Hosman, C. A., Hansen, B. B., and Holland, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Annals of Applied Statistics*, 4(2):849–870.

Hsu, J. Y., Roy, J. A., Xie, D., Yang, W., Shou, H., Anderson, A. H., Landis, J. R., Jepson, C., Wolf, M., Isakova, T., et al. (2017). Statistical methods for cohort studies of ckd: survival analysis in the setting of competing risks. *Clinical Journal of the American Society of Nephrology*, 12(7):1181–1189.

Hsu, J. Y. and Small, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics*, 69(4):803–811.

http://wonder.cdc.gov (2018). Centers for Disease Control and Prevention, National Center for Health Statistics. Compressed Mortality File on CDC WONDER Online Database, released December 2017. Data are from the Compressed Mortality File 1999-2016 Series 20 No. 2V, 2017, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed at http://wonder.cdc.gov on Mar 19, 2018.

Hudgens, M. G. and Halloran, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132.

Imbens, G. W. and Rosenbaum, P. R. (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 168(1):109–126.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Kang, H., Kreuels, B., Adjei, O., Krumkamp, R., May, J., and Small, D. S. (2013). The causal effect of malaria on stunting: a mendelian randomization and matching approach. *International Journal of Epidemiology*, 42(5):1390–1398.

Kang, H., Zhang, A., Cai, T. T., and Small, D. S. (2016). Instrumental variables estimation with some invalid instruments and its application to mendelian randomization. *Journal of the American Statistical Association*, 111(513):132–144.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research*, pages 141–167. Springer.

Keyfitz, N. (1966). Sampling variance of standardized mortality rates. *Human Biology*, 38(3):309–317.

Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 48(3):795–806.

Lachenbruch, P. A. (1998). Sensitivity, specificity, and vaccine efficacy. *Controlled clinical trials*, 19(6):569–574.

Lee, K. and Small, D. S. (2018). Estimating the malaria attributable fever fraction accounting for parasites being killed by fever and measurement error. *Journal of the American Statistical Association*, (just-accepted):1–38.

Lehmann, E. (1952). Testing multiparameter hypotheses. *The Annals of Mathematical Statistics*, pages 541–552.

Liu, W., Kuramoto, S. J., and Stuart, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science*, 14(6):570–580.

Mabunda, S., Aponte, J. J., Tiago, A., and Alonso, P. (2009). A country-wide malaria survey in mozambique. ii. malaria attributable proportion of fever and establishment of malaria case definition in children across different epidemiological settings. *Malaria Journal*, 8(1):74.

Mahmoudi, S. and Keshavarz, H. (2017). Efficacy of phase 3 trial of rts, s/as01 malaria vaccine: the need for an alternative development plan. *Human vaccines & immunotherapeutics*, 13(9):2098–2101.

Marcus, S. M. (1997). Using omitted variable bias to assess uncertainty in the estimation of an aids education treatment effect. *Journal of Educational and Behavioral Statistics*, 22(2):193–201.

Marsaglia, G. et al. (2006). Ratios of normal variables. *Journal of Statistical Software*, 16(4):1–10.

Martens, E. P., Pestman, W. R., de Boer, A., Belitser, S. V., and Klungel, O. H. (2006). Instrumental variables: application and limitations. *Epidemiology*, pages 260–267.

Martinussen, T., Vansteelandt, S., and Andersen, P. K. (2018). Subtleties in the interpretation of hazard ratios. *arXiv preprint arXiv:1810.09192*.

Matthews, K., Shepherd, J., and Sivarajasingham, V. (2006). Violence-related injury and the price of beer in england and wales. *Applied Economics*, 38(6):661–670.

McCandless, L. C., Gustafson, P., and Levy, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in medicine*, 26(11):2331–2347.

Meyer, B. D. (1995). Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, 13(2):151–161.

Moorthy, V., Reed, Z., and Smith, P. G. (2007). Measurement of malaria vaccine efficacy in phase iii trials: Report of a who consultation. *Vaccine*, 25(28):5115 – 5123.

Moorthy, V. S. and Ballou, W. R. (2009). Immunological mechanisms underlying protection mediated by rts, s: a review of the available data. *Malaria journal*, 8(1):312.

Morton, N. (2001). Linkage disequilibrium. In Brenner, S. and Miller, J. H., editors, *Encyclopedia of Genetics*, page 1105. Academic Press, New York.

Ndila, C. M., Uyoga, S., Macharia, A. W., Nyutu, G., Peshu, N., Ojal, J., Shebe, M., Awuondo, K. O., Mturi, N., Tsofa, B., et al. (2018). Human candidate gene polymorphisms and risk of severe malaria in children in kilifi, kenya: a case-control association study. *The Lancet Haematology*, 5(8):e333–e345.

Neuwald, A. F. and Green, P. (1994). Detecting patterns in protein sequences. *Journal of molecular biology*, 239(5):698–712.

Olotu, A., Fegan, G., Wambua, J., Nyangweso, G., Awuondo, K. O., Leach, A., Lievens, M., Leboulleux, D., Njuguna, P., Peshu, N., and et al. (2013). Four-year efficacy of rts,s/as01e and its interaction with malaria exposure. *New England Journal of Medicine*, 368(12):1111–1120.

Penny, M. A., Galactionova, K., Tarantino, M., Tanner, M., and Smith, T. A. (2015). The public health impact of malaria vaccine rts, s in malaria endemic africa: country-specific predictions using 18 month follow-up phase iii data and simulation models. *BMC medicine*, 13(1):170.

Penny, M. A., Maire, N., Studer, A., Schapira, A., and Smith, T. A. (2008). What should vaccine developers ask? simulation of the effectiveness of malaria vaccines. *PLoS One*, 3(9):e3193.

Rassen, J. A., Brookhart, M. A., Glynn, R. J., Mittleman, M. A., and Schneeweiss, S. (2009). Instrumental variables i: instrumental variables exploit natural variation in non-experimental data to estimate causal relationships. *Journal of clinical epidemiology*, 62(12):1226–1232.

Regules, J. A., Cummings, J. F., and Ockenhouse, C. F. (2011). The rts, s vaccine candidate for malaria. *Expert review of vaccines*, 10(5):589–599.

Reynolds, K. D. and West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Evaluation Review*, 11(6):691–714.

Rosenbaum, P. R. (1984). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, 79(387):565–574.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rosenbaum, P. R. (1992). Detecting bias with confidence in observational studies. *Biometrika*, 79(2):367–374.

Rosenbaum, P. R. (1999). Choice as an alternative to control in observational studies. *Statistical Science*, 14(3):259–278.

Rosenbaum, P. R. (2002a). Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association*, 97(457):183–192.

Rosenbaum, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327.

Rosenbaum, P. R. (2002c). *Observational studies*. Springer.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. Springer, New York.

Rosenbaum, P. R. (2011). Some approximate evidence factors in observational studies. *Journal of the American Statistical Association*, 106(493):285–295.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41.

Rosenbaum, P. R. and Silber, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, 104(488).

Rosenblum, M. and van der Laan, M. J. (2009). Using regression models to analyze randomized trials: Asymptotically valid hypothesis tests despite incorrectly specified models. *Biometrics*, 65(3):937–945.

Rosenblum, M. and van der Laan, M. J. (2010). Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. *The International Journal of Biostatistics*, 6(1).

Rubin, D. B. (1980). Comment (on d. basu, randomization analysis of experimental data: The fisher randomization test). *Journal of the American Statistical Association*, 75(371):591–593.

Rubin, D. B. (2005). Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

Rudolph, K. E., Stuart, E. A., Vernick, J. S., and Webster, D. W. (2015). Association between connecticut's permit-to-purchase handgun law and homicides. *American journal of public health*, 105(8):e49–e54.

Shaked, M. and Shanthikumar, J. G. (1994). *Stochastic Orders and Their Applications*. Academic Press, New York.

Shepherd, J. and Page, N. (2015). The economic downturn probably reduced violence far more than licensing restrictions. *Addiction*, 110(10):1583–1584.

Small, D. S., Cheng, J., and Ten Have, T. R. (2010). Evaluating the efficacy of a malaria vaccine. *The international journal of biostatistics*, 6(2).

Smith, G. D. and Ebrahim, S. (2008). Mendelian randomization: genetic variants as instruments for strengthening causal inference in observational studies. In *Biosocial Surveys*. National Academies Press (US).

Smith, T., Schellenberg, J. A., and Hayes, R. (1994). Attributable fraction estimates and case definitions for malaria in endemic. *Statistics in medicine*, 13(22):2345–2358.

Stanek, K. C., Iacono, W. G., and McGue, M. (2011). Returns to education: What do twin studies control? *Twin Research and Human Genetics*, 14(6):509–515.

Stensrud, M. J., Valberg, M., Røysland, K., and Aalen, O. O. (2017). Exploring selection bias by causal frailty models. *Epidemiology*, 28(3):379–386.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.*, 25(1):1–21.

Taylor, S. M., Parobek, C. M., and Fairhurst, R. M. (2012). Haemoglobinopathies and the clinical epidemiology of malaria: a systematic review and meta-analysis. *The Lancet infectious diseases*, 12(6):457–468.

Ter Kuile, F. O., Lal, A. A., Kariuki, S. K., Shi, Y. P., Mirel, L. B., Vulule, J. M., A., P.-H. P., Kolczak, M. S., Hawley, W. A., Nahlen, B. L., and et al. (2003). Impact of permethrin-treated bed nets on malaria, anemia, and growth in infants in an area of intense perennial malaria transmission in western kenya. *The American Journal of Tropical Medicine and Hygiene*, 68(4):68–77.

The RTS,S Clinical Trials Partnership (2011). First results of phase 3 trial of rts,s/as01 malaria vaccine in african children. *New England Journal of Medicine*, 365(20):1863–1875.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of computational and graphical statistics*, 12(1):156–175.

Tibshirani, R. and Redelmeier, D. A. (1997). Cellular telephones and motor-vehicle collisions: Some variations on matched-pairs analysis. *Canadian Journal of Statistics*, 25(4):581–591.

US Department of Health and Human Services (2004). Vital statistics of the united states: mortality, 1999 technical appendix. Technical report, CDC National Center for Health Statistics, Hyattsville, MD.

Van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data.* Springer Science & Business Media, New York, NY.

van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1).

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the E-value. *Annals of Internal Medicine*, 167(4):268–274.

VanderWeele, T. J. and Hernan, M. A. (2013). Causal inference under multiple versions of treatment. *Journal of Causal Inference*, 1(1):1–20.

Volpp, K. G., Rosen, A. K., Rosenbaum, P. R., Romano, P. S., Even-Shoshan, O., Wang, Y., Bellini, L., Behringer, T., and Silber, J. H. (2007). Mortality among hospitalized medicare beneficiaries in the first 2 years following ACGME resident duty hour reform. *Journal of the American Medical Association*, 298(9):975–983.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300.

Wang, L. and Krieger, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Statistics in medicine*, 25(13):2257–2271.

Webster, D., Crifasi, C. K., and Vernick, J. S. (2014). Effects of the repeal of missouris handgun purchaser licensing law on homicides. *Journal of Urban Health*, 91(2):293–302.

Wienke, A. (2010). Frailty models in survival analysis. *Chapman & Hall/CRC Biostatistics Series*.

Williams, T. N. (2011). How do hemoglobins s and c result in malaria protection? *Journal of Infectious Diseases*, 204(11):1651–1653.

Williams, T. N., Mwangi, T. W., Wambua, S., Alexander, N. D., Kortok, M., Snow, R. W., and Marsh, K. (2005). Sickle cell trait and the risk of plasmodium falciparum malaria and other childhood diseases. *The Journal of infectious diseases*, 192(1):178–186.

Wing, C., Simon, K., and Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health*, 39:453–469.

Wolf, A., Gray, R., and Fazel, S. (2014). Violence as a public health problem: An ecological study of 169 countries. *Social Science & Medicine*, 104:220–227.

Wolfe, D. A. (1974). A characterization of population weighted-symmetry and related results. *Journal of the American Statistical Association*, 69(347):819–822.

World Health Organization (2018). *Who world malaria report 2018*. World Health Organization, Luxembourg.

Wu, Y., Sinden, R. E., Churcher, T. S., Tsuboi, T., and Yusibov, V. (2015). Development of malaria transmission-blocking vaccines: from concept to product. In *Advances in parasitology*, volume 89, pages 109–152. Elsevier.

Yu, B. and Gastwirth, J. L. (2005). Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics*, 6(2):201–209.

Zubizarreta, J. R., Cerdá, M., and Rosenbaum, P. R. (2013). Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design. *Epidemiology*, 24(1):79–87.