



University of Pennsylvania
ScholarlyCommons

Publicly Accessible Penn Dissertations

2022

Removing Strong Data Assumptions In Causal Inference Via Large-Scale Optimization

Siyu Heng
University of Pennsylvania

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Applied Mathematics Commons](#), and the [Biostatistics Commons](#)

Recommended Citation

Heng, Siyu, "Removing Strong Data Assumptions In Causal Inference Via Large-Scale Optimization" (2022). *Publicly Accessible Penn Dissertations*. 5525.
<https://repository.upenn.edu/edissertations/5525>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/5525>
For more information, please contact repository@pobox.upenn.edu.

Removing Strong Data Assumptions In Causal Inference Via Large-Scale Optimization

Abstract

Many traditional and newly-developed causal inference approaches require imposing strong data assumptions, and if those assumptions were violated in practice, these approaches may be inapplicable, suffer from low statistical power, or lead to misleading causal conclusions. In this dissertation, we present three papers to show how large-scale optimization can sometimes aid in removing strong assumptions about the data generating process or the data collection procedure that are required by some existing causal inference approaches.

The first and second papers show how large-scale optimization can sometimes help remove strong assumptions about the data generating process. In the first paper, a new adaptive approach is proposed to combine two test statistics in matched observational studies. The proposed adaptive approach asymptotically uniformly dominates both of the two component test statistics in sensitivity analyses, regardless of the underlying data distribution. In the second paper, a model-free and finite-population-exact framework is proposed to analyze randomized experiments subject to outcome misclassification. This new framework is based on large-scale integer programming and can help researchers analyze a randomized experiment subject to outcome misclassification in a more comprehensive way without imposing any additional assumptions on a randomized experiment.

The third paper illustrates how large-scale optimization can help remove strong assumptions about the data collection procedure. Specifically, to study the effect of reducing malaria burden on the low birth weight rate in sub-Saharan Africa, a pair-of-pairs approach to a difference-in-differences study is proposed, which is built on optimal matching (a large-scale network flow problem) and cardinality matching (a large-scale integer programming problem). Unlike the traditional difference-in-differences studies, this pair-of-pairs approach does not require either panel data or repeated cross-sectional data to be collected before the analysis stage.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Applied Mathematics

First Advisor

Dylan S. Small

Subject Categories

Applied Mathematics | Biostatistics | Statistics and Probability

REMOVING STRONG DATA ASSUMPTIONS IN CAUSAL INFERENCE VIA
LARGE-SCALE OPTIMIZATION

Siyu Heng

A DISSERTATION

in

Applied Mathematics and Computational Science

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for

the Degree of Doctor of Philosophy

2022

Supervisor of Dissertation

Dylan S. Small, Universal Furniture Professor of Statistics and Data Science

Graduate Group Chairperson

Robin Pemantle, Merriam Term Professor of Mathematics

Dissertation Committee

Karun Adusumilli, Assistant Professor of Economics

Bhaswar B. Bhattacharya, Assistant Professor of Statistics and Data Science

Paul R. Rosenbaum, Robert G. Putzel Professor Emeritus of Statistics and Data Science

Dylan S. Small, Universal Furniture Professor of Statistics and Data Science

To my beloved father and mother

ACKNOWLEDGEMENT

First and foremost, I would like to show my deepest thanks to my academic advisor, Dylan Small, for his incredible support and warmest guidance throughout my Ph.D. study. I cannot imagine having a more nurturing and nicer Ph.D. advisor than him. When I was a first-year doctoral student starting working with Dylan, I had zero knowledge about causal inference and did not know much about statistics in general. Dylan still warmly accepted me as his Ph.D. student and met with me individually every week throughout the whole semester to teach me causal inference. After helping me get some causal inference knowledge, Dylan patiently and warmly guided me through several research projects that are all very significant for my research career, and meanwhile introduced many excellent collaborators to me throughout this process. Whenever I am frustrated by unexpected barriers in my research problems, Dylan always gives me priceless guidance, offers me valuable collaboration opportunities with other researchers, and encourages me in every single step. After helping me build up my own research taste through these research projects, Dylan encourages me to develop and work on my own research agenda, and meanwhile continuously offers his valuable advice and amazing support throughout the process. In one sentence, an epitome advisor is a scholar who does excellent research, and meanwhile is willing to take time and energy to do any big or small things for students' good. Dylan is such an advisor. The only way I can requite him is to treat my students as he treats me.

Next I would like to sincerely thank Karun Adusumilli, Bhaswar Bhattacharya, and Paul Rosenbaum for serving on my Ph.D. dissertation committee. To Paul, your books "*Observational Studies*" and "*Design of Observational Studies*" and your

academic papers have hugely influenced my causal inference research taste and agenda. Many of my research projects originate from and/or are inspired by your previous work. Especially, thank you for your warm guidance and support throughout our instrumental variable project with Dylan, which is my first research project on causal inference, and for your huge support and warm help throughout my job search process. To Karun and Bhaswar, thank you for all your kind help and warm advice on my statistics research throughout my Ph.D. study. Every single research meeting and discussion with you is enjoyable and inspiring.

I would like to show my sincere thanks to all my collaborators. I would like to express my gratitude to Wendy O'Meara and Pamela Shaw for their warm guidance and help throughout the research projects, as well as their huge support and help on my job search. I would like to thank Colin Fogarty, Hyunseung Kang, and Ryan Simmons for contributing heavily to the projects that compose the main contents of this dissertation. Many significant ideas and progress are owing to them and I learnt a lot from them throughout our collaboration experience. I would like to thank Bo Zhang, who is one of my closest student collaborators during my Ph.D. study. I learnt a lot about causal inference research from him. I would like to thank Ting Ye for several collaboration experiences and for her kind help and advice on my job talk and interviews. My thanks extend to Shuchi Anand, Kan Chen, Sameer Deshpande, Xu Han, Bikram Karmakar, Yuzhou Lin, Qi Long, Scott Lorch, Emily MacKay, and all the collaborators during my Ph.D. study for these wonderful collaboration experiences and many enjoyable academic discussions.

I am very grateful for all the faculty members and staff members from the Applied Mathematics and Computational Science (AMCS) Graduate Group and the Department of Statistics and Data Science. I would like to express my sincere thanks

to Charles Epstein, the inaugural AMCS Graduate Group Chair, for admitting me to the AMCS Graduate Group in 2016 and for all his warm support and guidance throughout my Ph.D. study. I am very grateful to Robin Pemantle, the current AMCS Graduate Group Chair, for all his guidance and help for me. I would like to express my gratitude to Jian Ding and Mark Low for offering extremely helpful advice and warm guidance on my job talk. My sincere thanks go to Ammarah Aftab for her kind support and warm guidance throughout my Ph.D. study and graduation process, and also to Gabriella Frisone, for her warm support during my job search process. I would like to thank all the faculty members who have taught me statistics, biostatistics, mathematics and English writing inside or outside the class, and all the staff members who have offered kind support and help to me during my study at the University of Pennsylvania.

I would like to thank my friends and peers from AMCS Graduate Group with whom I shared many fantastic memories and from whom I received a lot of kind help, including Zhiqi Bu, Kan Chen, Peiran Chen, Xingran Chen, Yansong Gao, Yebiao Jin, Yezheng Li, Xinyu Liao, Linbo Liu, Mingyang Liu, Lingxi Lu, Yifei Min, Yue Sheng, Xuran Wang, Jialin Yi, and many others. My thanks extend to classmates and office mates from the Department of Statistics and Data Science for many wonderful academic or non-academic discussions and experiences, including Shuxiao Chen, Sheng Gao, Hongming Pu, Hua Wang, Yichen Wang, Ruoqi Yu, and many others.

I would like to express my sincere thanks to all the faculty members and staff members who have offered help and support to me during my undergraduate study at Nanjing University and the University of Wisconsin-Madison, including Wei Cheng, David Sondak, Jean-Luc Thiffeault, Benedek Valko, and many others.

My sincere thanks extend to all my teachers during my study in Xuzhou, without whose guidance and help I was not able to have this opportunity of pursuing my Ph.D. study.

I am extremely grateful for the endless love and support from my parents. They are the people who love me most and whom I owe most on this earth. My deep gratitude extends to all my family members, especially my grandparents and younger brother, who have given so much love and support to me.

Finally, I would like to thank and praise God for his mercy and blessings upon me in every single step of my life.

ABSTRACT

REMOVING STRONG DATA ASSUMPTIONS IN CAUSAL INFERENCE VIA LARGE-SCALE OPTIMIZATION

Siyu Heng

Dylan S. Small

Many traditional and newly-developed causal inference approaches require imposing strong data assumptions, and if those assumptions were violated in practice, these approaches may be inapplicable, suffer from low statistical power, or lead to misleading causal conclusions. In this dissertation, we present three papers to show how large-scale optimization can sometimes aid in removing strong assumptions about the data generating process or the data collection procedure that are required by some existing causal inference approaches.

The first and second papers show how large-scale optimization can sometimes help remove strong assumptions about the data generating process. In the first paper, a new adaptive approach is proposed to combine two test statistics in matched observational studies. The proposed adaptive approach asymptotically uniformly dominates both of the two component test statistics in sensitivity analyses, regardless of the underlying data distribution. In the second paper, a model-free and finite-population-exact framework is proposed to analyze randomized experiments subject to outcome misclassification. This new framework is based on large-scale integer programming and can help researchers analyze a randomized experiment subject to outcome misclassification in a more comprehensive way without imposing any additional assumptions on a randomized experiment.

The third paper illustrates how large-scale optimization can help remove strong assumptions about the data collection procedure. Specifically, to study the effect of reducing malaria burden on the low birth weight rate in sub-Saharan Africa, a pair-of-pairs approach to a difference-in-differences study is proposed, which is built on optimal matching (a large-scale network flow problem) and cardinality matching (a large-scale integer programming problem). Unlike the traditional difference-in-differences studies, this pair-of-pairs approach does not require either panel data or repeated cross-sectional data to be collected before the analysis stage.

TABLE OF CONTENTS

| | |
|--|-----|
| ACKNOWLEDGEMENT | iii |
| ABSTRACT | vii |
| TABLE OF CONTENTS | ix |
| LIST OF TABLES | xi |
| LIST OF ILLUSTRATIONS | xvi |
| 1 Introduction | 1 |
| 2 Increasing Power for Observational Studies of Aberrant Response: An Adaptive Approach | 5 |
| 2.1 Introduction | 5 |
| 2.2 Notation and Reviews | 11 |
| 2.3 The Traditional Approach: the Mantel-Haenszel Test | 16 |
| 2.4 An Aberrant Rank Approach and Its Comparison with the Traditional Approach | 18 |
| 2.5 A New, General Adaptive Approach to Combine Two Test Statistics in Observational Studies | 28 |
| 2.6 Simulation Studies | 37 |
| 2.7 Adaptive Inference of the Effect of Mother's Age on Child Stunted Growth | 41 |
| 2.8 Discussion | 43 |
| 2.9 Appendices | 44 |

| | | |
|-----|---|-----|
| 3 | A Model-Free and Finite-Population-Exact Framework for Randomized Experiments Subject to Outcome Misclassification via Integer Programming | 88 |
| 3.1 | Introduction | 88 |
| 3.2 | Review | 95 |
| 3.3 | A Model-Free and Finite-Population-Exact Framework | 98 |
| 3.4 | Computing Warning Accuracy and Related Quantities | 106 |
| 3.5 | Real Data Application: Understanding the Puzzle in the PCPT | 119 |
| 3.6 | Summary | 123 |
| 3.7 | Appendices | 124 |
| 4 | Relationship Between Changing Malaria Burden and Low Birth Weight in Sub-Saharan Africa: A Difference-in-Differences Study via A Pair-of-Pairs Approach | 150 |
| 4.1 | Introduction | 150 |
| 4.2 | Materials and Methods | 153 |
| 4.3 | Results | 177 |
| 4.4 | Discussion | 189 |
| 4.5 | Appendices | 194 |
| | BIBLIOGRAPHY | 210 |

LIST OF TABLES

| | | |
|-----|---|----|
| 2.1 | Design sensitivities of the Mantel-Haenszel test and the aberrant rank test under Models 1-4 and matching with three controls with various parameters. The larger of the two design sensitivities of the two tests is in bold in each case. | 27 |
| 2.2 | Simulated power of the Mantel-Haenszel test, the aberrant rank test and the adaptive test. We set $\alpha = 0.05$, $c = 1$ and $m = 4$. We set $\beta = 1$ for Models 1 and 2 and $\delta = 2$ for Models 3 and 4. Each number is based on 2,000 replications. The largest of the three simulated powers in each case is in bold. | 38 |
| 2.3 | One-sided worst-case p-values under various Γ . The p-values ≈ 0.05 are in bold. We also report the approximate sensitivity value of each test with level 0.05. | 43 |
| 2.4 | Simulated power of T_1 , T_2 and the minimax procedure combining T_1 and T_2 | 62 |
| 2.5 | Simulated power of the minimax procedure with the various correlations of the two component test statistics. | 64 |
| 2.6 | Design sensitivities of the Mantel-Haenszel test and the aberrant rank test under Models 5-8 and matching with three controls with various parameters. The larger of the two design sensitivities of the two tests is in bold in each case. | 67 |

| | | |
|-----|--|-----|
| 2.7 | Simulated size of the Mantel-Haenszel test, the aberrant rank test and the adaptive test implementing Algorithm 1 with the above two tests as the component tests under the aberrant null. We set $\alpha = 0.05$, $c = 1$ and $m = 4$ (matching with three controls). | 79 |
| 2.8 | The total number of stunting cases among the treated individuals and controls in the matched data. | 82 |
| 3.1 | Illustration of the two types of symmetry: between-strata symmetry (e.g., stratum 1 and stratum 2) and within-strata symmetry (e.g., subject 1 and subject 2 in stratum 3). | 109 |
| 3.2 | The average computation time (in seconds), warning accuracy \mathcal{WA} and sensitivity weights $(W_T^{FP}, W_T^{FN}, W_C^{FP}, W_C^{FN})$ of different sets of $(E(N), p_0, p_1)$ for Simulation Scenarios 1 and 2 (for type I and type II randomization designs respectively). | 117 |
| 3.3 | The p-values, warning accuracy and sensitivity weights for the two binary outcomes of interest in the PCPT under Fisher's sharp null hypothesis of no treatment effect and alpha level 0.05. | 120 |
| 3.4 | Decomposition of the total number of outcome misclassification cases for each of the four types of outcome misclassification. | 121 |
| 3.5 | Simulations with Neyman's weak null. We report the average computation time (in seconds), warning accuracy \mathcal{WA} and sensitivity weights $(W_T^{FP}, W_T^{FN}, W_C^{FP}, W_C^{FN})$ of different sets of $(E(N), p_0, p_1)$ for Simulation Scenarios 1 and 3 (for type I and type II randomization designs respectively). | 148 |

| | | |
|-----|---|-----|
| 4.1 | The 19 selected sub-Saharan African countries along with their chosen early/late years of malaria prevalence (i.e., estimated parasite rate $PfPR_{2-10}$) and IPUMS-DHS early/late years. Note that some span over two successive years. | 157 |
| 4.2 | Summary of the Bayesian logistic regression model fitted over records with observed birth weight which is used to predict missing low birth weight indicators. | 170 |
| 4.3 | An interpretation of the coefficients of the intercept term and the three indicators defined in model (4.1) (i.e., the k_0, k_1, k_2, k_3) within each matched quadruple. The coefficient of the low malaria prevalence indicator (i.e., the k_1) incorporates the information of the magnitude of the effect of changing malaria burden (from high to low) on the low birth weight rate. | 173 |

4.4 The mean Haversine distance of the early year clusters and late year clusters is 24.1 km among the 219 high-low pairs of clusters, and 28.7 km among the 219 high-high pairs of clusters. The within-pair longitudes' and latitudes' correlations between the paired early year and late year clusters among the high-low and high-high pairs all nearly equal one. The mean values of the longitudes, the latitudes, the annual malaria prevalence (i.e., $PfPR_{2-10}$) measured at the early year, denoted as $PfPR_{2-10}^{Early}$, and at the late year, denoted as $PfPR_{2-10}^{Late}$, of the paired early year clusters (clusters sampled at the early year) and late year clusters (clusters sampled at the late year) among the 219 high-low and 219 high-high pairs of clusters used for the statistical inference respectively. Note that an early year cluster has a late year $PfPR_{2-10}$ and a late year cluster has an early year $PfPR_{2-10}$ since the MAP data contain $PfPR_{2-10}$ for each location and for each year between 2000 and 2015. 179

4.5 Balance of each covariate before matching (BM) and after matching (AM). We report the mean of each covariate (including early and late years) for high-low and high-high pairs of clusters, before and after matching. We also report each absolute standardized difference (Std.dif) before and after matching. 181

4.6 Inference with multiple imputation and mixed-effects linear probability model (4.1). The unit of estimates and CIs is a percentage point. 183

4.7 The early and late years coded in the IPUMS-DHS and GPS data sets. 195

| | | |
|------|---|-----|
| 4.8 | The numbers of the high-high pairs of clusters and high-low pairs of clusters contributed by each of the 19 selected sub-Saharan African countries after the matching in Step 1 and Step 2. We also summarize the total number of pairs of clusters after Step 1 matching in the first column. | 196 |
| 4.9 | Summary of the low malaria prevalence indicators, the time indicators, the group indicators, the covariates, and the birth weight records among the 18,112 study individual records. | 197 |
| 4.10 | Diagnostics for multiple imputation with the mixed-effects linear probability model. We report the between-imputation variance ("Between var"), the within-imputation variance ("Within var"), and the variance ratio: (between-imputation variance)/(within-imputation variance), denoted as "Var ratio". | 201 |
| 4.11 | The results of the sensitivity analyses for the coefficient of the low malaria prevalence indicator under various sensitivity parameters (p_1, p_2) divided into the four cases: Case 1: $p_1 > 0, p_2 > 0$; Case 2: $p_1 > 0, p_2 < 0$; Case 3: $p_1 < 0, p_2 > 0$; Case 4: $p_1 < 0, p_2 < 0$. The unit of estimates and CIs is a percentage point. | 209 |

LIST OF ILLUSTRATIONS

| | | |
|-----|--|-----|
| 2.1 | Covariate imbalances before and after matching with three controls. The plot reports the absolute standardized differences before and after matching of each covariate. The two dotted vertical lines are 0.1 and 0.2 cut-offs. | 85 |
| 4.1 | Work flow diagram of the study. | 163 |
| 4.2 | Formed quadruples (pairs of pairs) of matched high-low and high-high pairs of clusters. In Step 1, pairs of clusters from the early and late time periods are matched on geographic proximity and categorized as 'high-high' (comparison, or control) or 'high-low' (treated). In Step 2, pairs of high-high clusters are matched with pairs of high-low clusters based on cluster-level sociodemographic characteristics. The difference-in-differences estimate of the coefficient of changing malaria burden on the low birth weight rate is based on comparing (D–C) to (B–A). | 167 |

4.3 The estimated low birth weight rate of each cluster within the 219 high-high pairs and the 219 high-low pairs. The estimated low birth weight rate for each cluster are obtained from averaging over all the 500 imputed data sets of the 18,112 individual records. We draw a line to connect two paired clusters (one early year cluster and one late year cluster). Box plots for the low birth weight rates are also shown. Two of the four outliers of the late year clusters among the high-low pairs (i.e., the top four late year clusters in terms of low birth weight rate among the high-low pairs) may result from their extremely small within-cluster sample sizes (no more than 3 individual records for both two clusters). 185

1. Introduction

Causal inference seeks to study the causal effects of various human interventions and has deeply influenced decision-making in public health, biomedical research, economics, and social sciences. However, many traditional and newly-developed causal inference methods rely on strong assumptions about the data generating process and/or the data collection procedure, and if those assumptions do not hold, they may either be inapplicable or suffer from low power. Recent advances in efficient computation of large-scale optimization can aid in conducting valid and powerful causal inferences without imposing strong data assumptions. In this dissertation, we present three papers (corresponding to Chapter 2-4 respectively) to illustrate how large-scale optimization can sometimes help remove strong data assumptions in causal inference with experimental or observational data.

In Chapter 2 (the first paper), we illustrate how large-scale optimization can sometimes help remove strong assumptions about the data generating process in observational studies subject to unmeasured confounding. When studying the effect of a treatment on bad, aberrant outcomes, a traditional approach is to define a cutoff of a continuous score that is considered aberrant and use the Mantel-Haenszel test on the binary outcome of below the cutoff (or above the cutoff if the higher a response, the worse the outcome is). To make use of information about the severity of aberration (i.e., the amount a continuous score is below the cutoff) and improve power, Rosenbaum and Silber (2008, JASA) developed a new aberrant rank test. However, through proving a novel design sensitivity formula and related simulations, we show that although the aberrant rank test is indeed much more pow-

erful in a sensitivity analysis (i.e., more robust to unmeasured confounding) than the Mantel-Haenszel test for many data generating processes, there are also some cases in which the Mantel-Haenszel test is instead much more powerful. That is, how we choose between these two tests requires strong assumptions on the data generating process. To overcome this, we leverage large-scale optimization to develop a new, general adaptive approach, the two-stage programming method, and use it to adaptively combine the aberrant rank test and the Mantel-Haenszel test. We show our approach asymptotically dominates both tests and performs well in simulation studies, regardless of the unknown data generating process. We apply our approach to a study of the effect of teenage pregnancy on child stunting. This paper is joint work with Hyunseung Kang, Dylan Small, and Colin Fogarty, and was published in 2021 in Volume 83, Issue 3 of *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

In Chapter 3 (the second paper), we show how large-scale optimization can sometimes help remove strong assumptions about the data generating process in randomized experiments (trials) subject to outcome misclassification. Randomized experiments are the gold standard for making causal inferences as randomization can remove the need for assuming any data-generating (super-population) models. However, outcome misclassification (e.g., measurement error or response bias in binary outcomes) often exists in datasets and even a few misclassified outcomes may distort a causal conclusion drawn from a randomized experiment. All existing approaches to outcome misclassification rely on some data-generating model and therefore may not be applicable to randomized experiments without additional strong assumptions. We propose a model-free and finite-population-exact framework for randomized experiments subject to outcome misclassification,

which does not require adding any additional assumptions to a randomized experiment. A central quantity in our framework is “warning accuracy,” defined as the threshold such that the causal conclusion drawn from the measured outcomes may differ from that based on the true outcomes if the accuracy of the measured outcomes did not surpass that threshold. We discuss how learning the warning accuracy and related concepts can strengthen the design, analysis, and validation of a randomized experiment. We show that the warning accuracy can be computed efficiently (even for large-scale randomized experiments) by adaptively reformulating an integer program with respect to some intrinsic characteristic of various randomization designs. Our framework is applicable to both Fisher’s sharp null hypothesis and Neyman’s weak null hypothesis, covers a wide range of randomization designs, and can also be applied to matched/stratified observational studies adopting randomization-based inference. We apply our approach to analyze a large randomized clinical trial – the Prostate Cancer Prevention Trial. This paper is joint work with Pamela Shaw, of which a preliminary version was posted on *arxiv* (<https://arxiv.org/abs/2201.03111>).

In Chapter 4 (the third paper), through an applied work, we show how large-scale optimization can sometimes help remove strong assumptions on the data collection procedure for conducting causal inference with the difference-in-difference approach. Traditional difference-in-differences studies require either longitudinal data or repeated cross-sectional data. However, many large survey data sets sample different set of clusters at each time point, including the Demographic and Health Surveys (DHS), making traditional difference-in-differences inapplicable. To study the effect of a reduction in malaria prevalence on the low birth weight rate in sub-Saharan Africa using DHS data, we propose a novel study design—a

pair-of-pairs approach to a difference-in-differences study—to conduct a difference-in-differences study with neither longitudinal data nor repeated cross-sectional data. The proposed pair-of-pairs approach involves a two-step matching procedure, in which the first-step matching handles time and location heterogeneity through optimal matching (a large-scale discrete optimization problem) and the second-step matching involves cardinality matching (a large-scale integer programming problem). Using the proposed pair-of-pairs approach, we find that reducing malaria burden can potentially substantially reduce the low birth weight rate in sub-Saharan Africa, especially for first pregnancies. This paper is joint work with Wendy O’Meara, Ryan Simmons, and Dylan Small, which was published in 2021 in *eLife* (DOI: 10.7554/eLife.65133).

2. Increasing Power for Observational Studies of Aberrant Response: An Adaptive Approach

This chapter is adapted from "Heng, S., Kang, H., Small, D. S., and Fogarty, C. B. (2021). Increasing power for observational studies of aberrant response: an adaptive approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(3), 482-504."

2.1 Introduction

2.1.1 Examples of settings where interest is in the effect of treatment on aberrant response not average response

When evaluating the relative merits of competing treatment regimens, it can sometimes be more appropriate to focus on the effect of the treatment on poor outcomes (aberrant responses) rather than average outcomes. For example, malnutrition in children can cause both short- and long-term negative health outcomes and has been a long-standing global concern. According to the 2018 *Global Nutrition Report*, undernutrition contributes to around 45% of deaths among children under five. In studies on the effect of an exposure on child malnutrition, the most commonly used measurements of malnutrition are (1) stunting, (2) wasting, and (3) underweight. Stunting is defined as a child having a height less than or equal to

2 standard deviations below the mean height for the child's age (i.e., height-for-age z-score ≤ -2), where the mean and standard deviation come from a reference population such as the World Health Organization (WHO) Multicenter Growth Reference Study (WHO, 2006). Similarly, wasting and underweight are defined as weight-for-age z-score ≤ -2 and weight-for-height z-score ≤ -2 respectively; see WHO (1986), Harris et al. (2001) and Bloss et al. (2004). When studying causal determinants of malnutrition, say stunting, researchers typically focus on the pattern of stunting instead of the average treatment effect on the height of children. This is because being slightly below the average height will not cause any serious problems, but stunted growth can lead to adverse consequences for the child including poor cognition and educational performance, low adult wages and lost productivity (WHO, 2017). The standard approach in studies of causal determinants of malnutrition is to consider stunting, wasting or underweight as binary outcomes, and to test the null that the treatment (potential causal determinant) does not affect that binary outcome for each individual through either Fisher's exact test for unstratified data or the Mantel-Haenszel test with stratified data (e.g., Brown et al., 1982; Walker et al., 1991; Bloss et al., 2004; Garrett and Ruel, 2005; Phuka et al., 2008; Null et al., 2018).

Numerous causal problems share a similar structure with that of the causal determinants of malnutrition, where we care about whether a certain treatment would change the pattern of some aberrant response (e.g., stunted growth) rather than the average treatment effect over the whole population; see Appendix G in the supplementary materials for more examples. Rosenbaum and Silber (2008) referred to this as the aberrant effects of treatment problem. When studying aberrant effects, researchers typically choose a widely-used cut-off to define a dichotomous

outcome (e.g., stunted or not; wasted or not) from a continuous response (e.g., height-for-age z-score; weight-for-age z-score), and then perform Fisher’s exact test or the Mantel-Haenszel test. These traditional methods are both simple and convenient, but discard potentially useful information on the severity of aberrant response (e.g., exact height-for-age z-scores of children with stunted growth) and thus may fail to detect existing aberrant effects of the treatment.

2.1.1.1 A matched observational study on the effect of teenage pregnancy on stunting

Prior work has suggested that teen mothers are more likely to bear stunted children and some studies have tried to investigate whether teenage pregnancy has a causal effect on child stunting (e.g., [Van de Poel et al., 2007](#); [Darteh et al., 2014](#)). We examine this causal question using data from the Kenya 2003 Demographic and Health Surveys (DHS). Using Kenya’s definition of adulthood, we define children with mother’s age ≤ 18 years as treated individuals, and children with mother’s age ≥ 19 years as controls. We use their height-for-age z-scores as the outcomes, where the z-score is with respect to the WHO Multicenter Reference Growth Study ([WHO, 2006](#)). Recall that according to the WHO, low child height-for-age, or “stunting,” is defined as height-for-age z-score ≤ -2 . We conduct a matched observational study where we match each treated individual with controls on seven covariates: mother’s highest education level; geographic district; household wealth index in quantiles; household’s main source of drinking water; household’s toilet facilities; sex; and children’s age in years. Matching is a transparent and easily understandable way of adjusting for observed covariates and has been widely applied in observational studies ([Rosenbaum, 2002b](#); [Hansen, 2004](#); [Stuart, 2010](#); [Zubizarreta, 2012](#); [Pimentel et al., 2015](#)). We discarded 1466 records

with missing or unspecified height-for-age z-scores, source of drinking water or toilet facilities, leaving 4483 records. Among these 4483 children, there are 150 treated individuals and we matched each to three controls, 450 controls in total. We used optimal matching using rank-based Mahalanobis distance with a propensity score caliper (Hansen and Klopfer, 2006). To evaluate the balance on baseline covariates before and after matching, we use standardized differences which are defined as a weighted difference in means divided by the pooled standard deviation between the treated and control groups before matching; see Chapter 9 of Rosenbaum (2010). The standardized differences of the seven covariates are all near zero after matching, indicating good balance; see Appendix I for details.

2.1.2 Our Contributions

Previous work on inference for aberrant effects of treatment has considered randomized trials where there is no unmeasured confounding by design (Rosenbaum and Silber, 2008). In an observational study, we typically worry about unmeasured confounding and would like to have an approach that has good power to detect an effect that is insensitive to a moderate amount of unmeasured confounding (Rosenbaum, 2004, 2010). In this paper, we develop an adaptive approach for inference about aberrant treatment effects from matched observational studies that asymptotically uniformly dominates the traditional approach of performing the Mantel-Haenszel test based on a dichotomous outcome of aberrant/not aberrant.

Our new approach is developed in two parts. In the first part, we introduce the aberrant null hypothesis of no treatment effect for matched studies which is especially suitable for studying aberrant treatment effects, and then introduce the aberrant rank test for matched studies along with its sensitivity analysis and study

its asymptotic power. The aberrant rank test takes the form of the sum of aberrant ranks among all the treated units, with the Wilcoxon rank sum test as a special case. It is more powerful than the Mantel-Haenszel test for testing the aberrant null of no treatment effect in many settings because it considers not only the incidence of aberrant response, but also the severity of aberration. We formally demonstrate this through proving a novel design sensitivity formula. Design sensitivity measures the limiting robustness of a test to hidden bias in an observational study as the sample size increases: larger design sensitivity corresponds to greater asymptotic robustness to hidden bias (Rosenbaum, 2004, 2010). Our new design sensitivity formula allows us to asymptotically compare the performances of the aberrant rank test and the Mantel-Haenszel test for testing the aberrant null under various settings. We also validate that our asymptotic findings provide good guidance for realistic sample sizes in simulation studies. We illustrate that whether we should use the aberrant rank test or the Mantel-Haenszel test depends on the unknown data generating process, and making the wrong choice can substantially harm the performance of a sensitivity analysis. The proof of the design sensitivity formula involves a new technique that uses empirical processes to analyze the asymptotics of matched observational studies as the number of matched strata grows and can be of independent interest.

In the second part, we develop a novel, general adaptive approach called "the two-stage programming method" to combine two tests in observational studies such that the design sensitivity of the resulting adaptive test is always greater than or equal to maximum of the design sensitivities of the two component tests performed in isolation, regardless of the underlying data generating distribution. We refer to this newly discovered phenomenon as "super-adaptivity." Thus, applying

our new adaptive approach to combine the aberrant rank test and the Mantel-Haenszel test uniformly dominates the traditional approach based solely upon the Mantel-Haenszel test in terms of the design sensitivity. We evaluate our adaptive test via simulations and show that under various settings its power is close to the maximal power of its components (i.e., the aberrant rank test and the Mantel-Haenszel test) for realistic sample sizes, and therefore avoids potentially drastic reductions in power stemming from making the wrong choice between the two component tests.

The first adaptive approach for sensitivity analysis was introduced in [Rosenbaum \(2012\)](#). It has been applied to or modified for various settings (e.g., [Zubizarreta et al., 2014](#); [Rosenbaum and Small, 2017](#); [Ertefaie et al., 2018](#); [Zhao et al., 2018](#); [Shauly-Aharonov, 2020](#)). Our new adaptive approach goes beyond this traditional adaptive approach in two aspects. First, the traditional adaptive approach only works if all of the component test statistics are stochastically dominated by a known distribution in a matched observational study. While commonly used tests statistics for binary outcomes or general outcomes in pair-matched studies have this property, most commonly used test statistics for general outcomes in matched studies that allow for multiple controls or full matching do not have this stochastic dominance property, including the aberrant rank test, the Wilcoxon rank sum test, the Hodges-Lehmann aligned rank test and the Huber-Maritz m-tests ([Gastwirth et al., 2000](#); [Rosenbaum, 2002b, 2007](#)). In contrast, our new adaptive approach covers most of the existing testing scenarios in matched studies as it works for any sum statistics and various matching techniques, including pair matching, matching with multiple controls and full matching. Second, our new adaptive approach uniformly dominates the traditional approach in terms of design sensitivity, which

is an immediate consequence of the super-adaptivity property.

2.2 Notation and Reviews

2.2.1 Matching-based randomization inference and sensitivity analysis

Suppose there are I strata $i = 1, \dots, I$. Each stratum contains m ($m \geq 2$) individuals (e.g., children) where one individual received treatment and the other $m - 1$ individuals received control. Let $Z_{ij} = 1$ if individual j in stratum i received treatment (e.g., mother's age ≤ 18 years), otherwise let $Z_{ij} = 0$ (e.g., mother's age ≥ 19 years). Denote the collection of treatment assignments as $\mathbf{Z} = (Z_{11}, \dots, Z_{Im})^T$. Let \mathcal{Z} be the set of all possible values of \mathbf{Z} where $\mathbf{Z} \in \mathcal{Z}$ if and only if $\sum_{j=1}^m Z_{ij} = 1$ for all i . Let $|S|$ denote the number of elements of a finite set S , then we have $|\mathcal{Z}| = m^I$. Let \mathbf{x}_{ij} and u_{ij} denote the observed covariates and an unobserved covariate respectively for each individual j in stratum i . Typically, each stratum i is formed by matching on the observed covariates such that $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$ or $\mathbf{x}_{ij} \approx \mathbf{x}_{ij'}$. However, matching cannot directly adjust for the unobserved covariate, so $u_{ij} \neq u_{ij'}$ is possible. Under the potential outcomes framework, if individual j in stratum i received treatment ($Z_{ij} = 1$), we observe the potential response r_{Tij} ; otherwise ($Z_{ij} = 0$), we observe r_{Cij} . That is, the observed response (e.g., the observed height-for-age z-score) for individual ij is $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$ (Neyman, 1923; Rubin, 1974). Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, \dots, I, j = 1, \dots, m\}$. Denote the collection of responses as $\mathbf{R} = (R_{11}, \dots, R_{Im})^T$. Fisher's sharp null hypothesis of no treatment effect asserts that $H_0 : r_{Tij} = r_{Cij}, \forall i, j$.

In a randomized experiment, where we can assume random treatment assignment

in each stratum, i.e., $\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\mathcal{Z}| = m^{-I}$ for all $\mathbf{z} \in \mathcal{Z}$, the significance level of a test statistic T being greater than or equal to the observed value t under the null hypothesis can be calculated via randomization inference:

$$\begin{aligned} \mathbb{P}(T \geq t \mid \mathcal{F}, \mathcal{Z}) &= \sum_{\mathbf{z} \in \mathcal{Z}} \mathbb{1}\{T(\mathbf{z}, \mathbf{R}) \geq t\} \cdot \mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) \\ &= \frac{|\{\mathbf{z} \in \mathcal{Z} : T(\mathbf{z}, \mathbf{R}) \geq t\}|}{|\mathcal{Z}|}, \end{aligned} \quad (2.1)$$

where $\mathbb{1}(A) = 1$ if A is true, and $\mathbb{1}(A) = 0$ otherwise. For large I , (2.1) can be approximated via asymptotic normality of the null distribution of T (Rosenbaum, 2002b).

In an observational study, however, it is unrealistic to assume that the treatment is randomly assigned in each stratum even if we have matched on all the observed covariates, due to the possible presence of unobserved covariates (unmeasured confounders). A sensitivity analysis tries to determine how departures from random assignment of treatment would affect inferences on treatment effects. Let $\pi_{ij} = \mathbb{P}(Z_{ij} = 1 \mid \mathcal{F})$ denote the probability that, in the population before matching, individual ij will receive treatment. We follow the widely-used Rosenbaum sensitivity analysis framework (Rosenbaum, 2002b) which considers a logit model linking π_{ij} to \mathbf{x}_{ij} and normalized u_{ij} :

$$\log \left(\frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \theta(\mathbf{x}_{ij}) + \gamma u_{ij}, \quad \text{where } u_{ij} \in [0, 1], \quad (2.2)$$

where $\theta(\mathbf{x}_{ij})$ is an arbitrary unknown function of \mathbf{x}_{ij} and $\gamma \geq 0$ is a sensitivity parameter. The assumption $u_{ij} \in [0, 1]$ is no more restrictive than assuming a bounded support of u_{ij} and is only imposed to make γ more interpretable (Rosen-

baum, 2002b). Under (2.2), if we imagine that each (r_{Tij}, r_{Cij}) is drawn from some super-population model (only for interpretation, not necessary for randomization inference) and is confounded by the set of covariates $(\mathbf{x}_{ij}, u_{ij})$, then the strong ignorability assumption (Rosenbaum and Rubin, 1983a) would hold were $(\mathbf{x}_{ij}, u_{ij})$ to all be measured (but in fact u_{ij} is not measured), i.e., we have $(r_{Tij}, r_{Cij}) \perp\!\!\!\perp Z_{ij} \mid \mathbf{x}_{ij}, u_{ij}$ and $0 < \mathbb{P}(Z_{ij} = 1 \mid \mathbf{x}_{ij}, u_{ij}) < 1$ for all i, j . Here u_{ij} can also represent an aggregate measurement of all potential, perhaps more than one, unmeasured confounders u_{ij1}, u_{ij2}, \dots . For example, if $\log\{\pi_{ij}/(1 - \pi_{ij})\} = \theta(\mathbf{x}_{ij}) + g(u_{ij1}, u_{ij2}, \dots)$ for some function g with bounded support $[0, \zeta]$, then (2.2) holds with setting $u_{ij} = \zeta^{-1}g(u_{ij1}, u_{ij2}, \dots)$ and $\gamma = \zeta$.

Under model (2.2), it is straightforward to show that for any two individuals ij and ij' within the same stratum, the ratio of their odds of receiving the treatment is bounded by the sensitivity parameter $\Gamma = \exp(\gamma) \geq 1$:

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1 - \pi_{ij'})}{\pi_{ij'}(1 - \pi_{ij})} \leq \Gamma, \quad \text{for all } i, j, j' \text{ with } \mathbf{x}_{ij} = \mathbf{x}_{ij'}.$$

Note that $\Gamma = 1$ is equivalent to random assignment. The more Γ departs from 1, the more the treatment assignment potentially departs from random assignment. It is then easy to show that model (2.2) implies the following biased treatment assignment probability after matching, assuming $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$:

$$\mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = \prod_{i=1}^I \frac{\exp(\gamma \sum_{j=1}^m z_{ij} u_{ij})}{\sum_{j=1}^m \exp(\gamma u_{ij})}, \quad \mathbf{z} \in \mathcal{Z}, 0 \leq u_{ij} \leq 1.$$

Then researchers usually look at *the worst-case p-value*, which is defined as the largest p-value given the sensitivity parameter Γ over all possible arrangements of unmeasured confounders u_{ij} . For example, for a one-sided test, the

worst-case p-value reported by a test statistic T given its observed value t with the sensitivity parameter $\Gamma = \exp(\gamma)$ is $\max_{0 \leq u_{ij} \leq 1} \mathbb{P}(T \geq t \mid \mathcal{F}, \mathcal{Z}) = \max_{0 \leq u_{ij} \leq 1} \sum_{\mathbf{z} \in \mathcal{Z}} \mathbb{1}\{T(\mathbf{z}, \mathbf{R}) \geq t\} \cdot \mathbb{P}(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z})$. In practice, researchers gradually increase Γ , compute the worst-case p-value for each Γ , and report the *sensitivity value*, which is defined as the largest Γ such that the corresponding worst-case p-value exceeds some prespecified level α and informs the magnitude of hidden bias required to alter the causal conclusion drawn from the primary analysis assuming no unmeasured confounding (Zhao, 2018). For other models of sensitivity analysis, see Shepherd et al. (2006), McCandless et al. (2007), Mitra and Heitjan (2007), Hosman et al. (2010), Keele and Quinn (2017), VanderWeele and Peng (2007), and Zhao (2018).

2.2.2 Power of a sensitivity analysis and design sensitivity

The power of a test is the probability that the test will successfully reject the null hypothesis and is calculated under some alternative. In parallel, the power of a sensitivity analysis is the probability that the test will correctly reject the null, for any possible arrangements of unmeasured confounders given some $\Gamma \geq 1$, under some alternative. To be more specific, for a fixed Γ , the power of an α level sensitivity analysis using a test statistic T is calculated as the probability that the worst-case p-value corresponding to T falls below α when conducting a sensitivity analysis at Γ . When calculating the power, we need to specify a data generating process for the alternative. Following previous work, we consider power under the alternative, specifically, a “favorable situation” where there is a treatment effect of a specified magnitude and no hidden bias (Rosenbaum, 2004, 2010). Even though there is no hidden bias in this favorable situation, we would typically not know that for sure in an observational study and would prefer a test statistic with a

higher power of sensitivity analysis for plausible values of $\Gamma \geq 1$ (i.e., with high degree of insensitivity to hidden bias). This strategy of calculating the power is more appropriate than those assuming alternatives of both a treatment effect and a bias in treatment assignment. For example, suppose that we instead use the alternative of a small treatment effect and a large bias in treatment assignment, then a test statistic that has a high chance of rejecting the null hypothesis of no treatment effect (i.e., high statistical power) with small or moderate Γ may not be favorable because we cannot tell if its high statistical power results from its detection of the actual treatment effect or the underestimation of the magnitude of hidden bias. This ambiguity will not occur when calculating the power under the favorable situation in which the researchers always seek to reject the null under plausible Γ . Therefore, calculating the power of a sensitivity analysis under the favorable situation provides a logically consistent way to compare competing test statistics in observational studies. See [Hansen et al. \(2014\)](#) and [Rosenbaum \(2017\)](#) (Chapter 10) for detailed discussion on this.

Typically, under some regularity assumptions (varying with different matching structures and tests) on the data generating process of responses \mathbf{R} , there is a number $\tilde{\Gamma}$ called the design sensitivity, such that as the sample size $I \rightarrow \infty$, the power of a sensitivity analysis goes to 1 if the analysis is performed with $\Gamma < \tilde{\Gamma}$, and the power goes to 0 if performed with $\Gamma > \tilde{\Gamma}$. That is, $\tilde{\Gamma}$ refers to the sharp transition of consistency of a test in a sensitivity analysis ([Rosenbaum, 2004](#)). The design sensitivity gives us a powerful and elegant tool to asymptotically compare two test statistics or two study designs under each data distribution model - the test or the study design with a larger $\tilde{\Gamma}$ is asymptotically more robust to unmeasured confounders. Besides its mathematical elegance, the design sensitivity has

been shown to be a powerful tool in practical studies (Stuart and Hanna, 2013; Zubizarreta et al., 2013).

2.3 The Traditional Approach: the Mantel-Haenszel Test

In settings such as those described in Section 2.1.1, there is a subset $\mathcal{A} \subset \mathbb{R}$ such that any $R_{ij} \in \mathcal{A}$ is considered as an aberrant response, and researchers care about whether the treatment would change the pattern of aberrant response instead of the average treatment effect. In these settings, a typical approach is to define a dichotomous outcome $\tilde{R}_{ij} = \mathbb{1}(R_{ij} \in \mathcal{A})$ indicating whether individual ij had an aberrant outcome or not. For example, in Section 2.1.1.1 where we focus on whether child ij showed stunted growth (i.e., $R_{ij} \leq -2$), we can let $\mathcal{A} = (-\infty, -2]$ and the dichotomous observed outcome indicating stunted growth $\tilde{R}_{ij} = \mathbb{1}(R_{ij} \leq -2)$ is binary. Let $\tilde{r}_{Tij} = \mathbb{1}(r_{Tij} \in \mathcal{A})$ and $\tilde{r}_{Cij} = \mathbb{1}(r_{Cij} \in \mathcal{A})$, we have $\tilde{R}_{ij} = \mathbb{1}(R_{ij} \in \mathcal{A}) = Z_{ij}\mathbb{1}(r_{Tij} \in \mathcal{A}) + (1 - Z_{ij})\mathbb{1}(r_{Cij} \in \mathcal{A}) = Z_{ij}\tilde{r}_{Tij} + (1 - Z_{ij})\tilde{r}_{Cij}$. Then researchers focus on a categorized Fisher's sharp null of no treatment effect $\tilde{H}_0 : \tilde{r}_{Tij} = \tilde{r}_{Cij}, \forall i, j$, that is, whether individual ij would show aberrant response or not will not be affected by whether he or she received the treatment or not. Note that \tilde{H}_0 does not imply any information about the severity of aberration. It is clear that if H_0 holds true, so does \tilde{H}_0 , and if \tilde{H}_0 is false, so is H_0 .

The traditional approach then performs the Mantel-Haenszel test (Mantel and Haenszel, 1959), which can be regarded as an analogue of Fisher's exact test when there are two or more stratum, $I \geq 2$. Formally, the Mantel-Haenszel test utilizes the statistic $T_{M-H} = \sum_{i=1}^I \sum_{j=1}^m Z_{ij}\mathbb{1}(R_{ij} \in \mathcal{A}) = \sum_{i=1}^I \sum_{j=1}^m Z_{ij}\tilde{R}_{ij}$, which is the

number of aberrant responses among treated individuals. For matched pairs, the Mantel-Haenszel test reduces to McNemar's test (Cox and Snell, 2018). In a randomized experiment, we can use (2.1) to conduct randomization inference. In an observational study, we can use the related result in Rosenbaum (2002b) (Chapter 4) to conduct sensitivity analyses: under matching with $m - 1$ controls, for any t , the one-sided worst-case p-value $\max_{0 \leq u_{ij} \leq 1} \mathbb{P}(T \geq t \mid \mathcal{F}, \mathcal{Z}) = \mathbb{P}(T^+ \geq t \mid \mathcal{F}, \mathcal{Z})$ where T^+ is the sum of I independent Bernoulli random variables B_1, \dots, B_I , with B_i taking value one with probability $p_i^+ = \{\Gamma \sum_{j=1}^m \mathbb{1}(R_{ij} \in \mathcal{A})\} / \{(\Gamma - 1) \sum_{j=1}^m \mathbb{1}(R_{ij} \in \mathcal{A}) + m\}$ and value zero with probability $1 - p_i^+$. Therefore, we have as $I \rightarrow \infty$, for each fixed t ,

$$\max_{0 \leq u_{ij} \leq 1} \mathbb{P}(T \geq t \mid \mathcal{F}, \mathcal{Z}) = \mathbb{P}(T^+ \geq t \mid \mathcal{F}, \mathcal{Z}) \simeq 1 - \Phi \left(\frac{t - \sum_{i=1}^I p_i^+}{\sqrt{\sum_{i=1}^I p_i^+ (1 - p_i^+)}} \right), \quad (2.3)$$

where Φ is the distribution function of standard normal distribution and ' \simeq ' denotes that two sequences are asymptotically equal. We can then use (2.3) to report worst-case p-values for various sensitivity parameters Γ . The design sensitivity of the Mantel-Haenszel test has also been derived in Rosenbaum and Small (2017).

The Mantel-Haenszel test is simple and convenient but can lose power from ignoring information about the magnitude of aberration.

2.4 An Aberrant Rank Approach and Its Comparison with the Traditional Approach

2.4.1 The aberrant null and the aberrant rank test

Although H_0 and \tilde{H}_0 are widely used null hypotheses in randomized experiments and observational studies, they do not best capture the hypotheses of interest in studying the causal determinants of aberrant response when severity of aberration matters. To better capture the hypothesis of interest, [Rosenbaum and Silber \(2008\)](#) introduced the aberrant null hypothesis of no effect of treatment on individuals who would have an aberrant response under either the treatment or control. Formally, as in Section 2.3, let \mathcal{A} be a subset of \mathbb{R} that defines an aberrant response. Then the null hypothesis of no aberrant effect states that

$$H_0^A : r_{Tij} = r_{Cij}, \forall i, j, \text{ if either } r_{Tij} \in \mathcal{A} \text{ or } r_{Cij} \in \mathcal{A}.$$

It is easy to see that H_0^A is a weaker hypothesis than Fisher's sharp null H_0 , in the sense that H_0 implies H_0^A , but the converse is not true. And we can also see that H_0^A is a stronger hypothesis than the categorized Fisher's sharp null \tilde{H}_0 , in the sense that H_0^A implies \tilde{H}_0 , but the converse is not true. That is, H_0^A is a null hypothesis that lies between H_0 and \tilde{H}_0 .

Let us consider studying a potential causal determinant of stunting to illustrate why H_0^A is a more appropriate null hypothesis to test when the pattern of aberration is our main focus. Note that all the alternatives can be classified into the following four cases: (i) **Case 1:** $r_{Tij} \in \mathcal{A}, r_{Cij} \notin \mathcal{A}$, i.e., treatment will cause stunting for child ij ; (ii) **Case 2:** $r_{Tij} \notin \mathcal{A}, r_{Cij} \in \mathcal{A}$, i.e., treatment will prevent stunting

for child ij ; (iii) **Case 3**: $r_{Tij} \in \mathcal{A}$, $r_{Cij} \in \mathcal{A}$, and $r_{Tij} \neq r_{Cij}$, i.e., treatment will not prevent stunting for child ij , but it will affect the severity of stunting; (iv) **Case 4**: $r_{Tij} \notin \mathcal{A}$, $r_{Cij} \notin \mathcal{A}$, and $r_{Tij} \neq r_{Cij}$, i.e., child ij will not show stunted growth no matter whether he or she received treatment or not. Thus, H_0^A is against Cases 1-3, while H_0 is against all the four cases and \tilde{H}_0 is against only Cases 1 and 2. Our goal is to decide whether the treatment affects stunted growth. It is clear that in Cases 1 and 2, the treatment affects stunted growth (causing stunting in Case 1, preventing stunting in Case 2). Case 3 also indicates the treatment affects stunted growth, since although the treatment will not prevent a child from being stunted, it will affect the severity of stunting, i.e., it will aggravate or alleviate the child's stunting growth which could have a huge impact on the child. In Case 4, the treatment does not affect stunted growth since the child will be healthy and non-stunted no matter whether he or she is exposed to the treatment or control. Consideration of these four cases shows that H_0^A is a more appropriate null hypothesis than H_0 and \tilde{H}_0 because it contains in the alternative the three cases where treatment affects stunted growth but keeps in the null the fourth case where treatment does not affect stunted growth.

In this paper, our argument focuses on \mathcal{A} with the form of $\mathcal{A} = [c, +\infty)$ (or equivalently, $(c, +\infty)$) for some $c \in \mathbb{R}$. In these settings, there is a threshold value c indicating aberration, which is common in practical research. The argument works in parallel with $\mathcal{A} = (-\infty, c]$ and $\mathcal{A} = (-\infty, c)$. For example, according to the WHO, stunting is defined as height-for-age z-score ≤ -2 , and in this case $\mathcal{A} = (-\infty, c]$ with $c = -2$.

Rosenbaum and Silber (2008) introduced the aberrant rank test for randomized experiments with unmatched data, and Small et al. (2013) considered the aberrant

rank in case-referent studies. In this paper, we derive a new aberrant rank test for matched observational cohort studies. We define $q(v \mid \mathbf{R}) = \sum_{i'=1}^I \sum_{j'=1}^m \mathbb{1}(v \geq R_{ij'} \geq c)$ and refer to $q(R_{ij} \mid \mathbf{R})$ as the aberrant rank of individual ij . There are some features worth mentioning. First, the aberrant rank $q(R_{ij} \mid \mathbf{R})$ depends on all the responses, including those that are not in the same stratum as R_{ij} . Second, if individual ij did not show aberrant response, $q(R_{ij} \mid \mathbf{R})$ is zero, and if he or she did show aberrant response, $q(R_{ij} \mid \mathbf{R})$ takes the rank of R_{ij} among all the responses of individuals with aberrant response. Third, $q(v \mid \mathbf{R})$ is monotonic in v .

Next, we define the aberrant rank test for a stratified (e.g., matched) study as

$$T_{\text{abe}} = \sum_{i=1}^I \sum_{j=1}^m Z_{ij} q(R_{ij} \mid \mathbf{R}), \quad (2.4)$$

which is the sum of all the aberrant ranks over all treated individuals. When $c = -\infty$, T_{abe} reduces to the Wilcoxon rank sum test. Under the null hypothesis of no aberrant effects H_0^A , $q(R_{ij} \mid \mathbf{R})$ is fixed. In a randomized experiment, we can use (2.1) along with its asymptotic approximation to report p-values. In a sensitivity analysis, unlike the Mantel-Haenszel test, in general we cannot find a known distribution to bound the distribution of T_{abe} . However, under pair matching or matching with multiple controls, utilizing the asymptotic separability algorithm in [Gastwirth et al. \(2000\)](#), for any given t , we can approximate the one-sided worst-case p-value $\max_{0 \leq u_{ij} \leq 1} \mathbb{P}(T_{\text{abe}} \geq t \mid \mathcal{F}, \mathcal{Z})$ under H_0^A . Let $b \in \{1, \dots, m-1\} =: [m-1]$, and $\bar{\mu}_{ib}$ and \bar{v}_{ib} be the null expected value and variance of $\sum_{j=1}^m Z_{ij} q(R_{ij} \mid \mathbf{R})$ under $u_{i1} = \dots = u_{ib} = 0$ and $u_{i,b+1} = \dots = u_{im} = 1$ with different values of b respectively: for $i = 1, \dots, I$ and $b = 1, \dots, m-1$, we

have

$$\begin{aligned}\bar{\bar{\mu}}_{ib} &= \frac{\sum_{j=1}^b q(R_{i(j)} | \mathbf{R}) + \Gamma \sum_{j=b+1}^m q(R_{i(j)} | \mathbf{R})}{b + \Gamma(m - b)}, \\ \bar{\bar{v}}_{ib} &= \frac{\sum_{j=1}^b q^2(R_{i(j)} | \mathbf{R}) + \Gamma \sum_{j=b+1}^m q^2(R_{i(j)} | \mathbf{R})}{b + \Gamma(m - b)} - \bar{\bar{\mu}}_{ib}^2,\end{aligned}$$

where we rearrange R_{i1}, \dots, R_{im} as $R_{i(1)} \leq \dots \leq R_{i(m)}$. Let $\bar{\bar{\mu}}_i = \max_{b \in [m-1]} \bar{\bar{\mu}}_{ib}$, $B_i = \{b : \bar{\bar{\mu}}_{ib} = \bar{\bar{\mu}}_i, b \in [m-1]\}$ and $\bar{\bar{v}}_i = \max_{b \in B_i} \bar{\bar{v}}_{ib}$. Then the one-sided worst-case p-value can be approximated via

$$\max_{0 \leq u_{ij} \leq 1} \mathbb{P}(T_{\text{abe}} \geq t | \mathcal{F}, \mathcal{Z}) \simeq 1 - \Phi\left(\frac{t - \sum_{i=1}^I \bar{\bar{\mu}}_i}{\sqrt{\sum_{i=1}^I \bar{\bar{v}}_i}}\right) \text{ as } I \rightarrow \infty.$$

Letting $\zeta_\alpha = \sum_{i=1}^I \bar{\bar{\mu}}_i + \Phi^{-1}(1 - \alpha) \sqrt{\sum_{i=1}^I \bar{\bar{v}}_i}$, $\Psi_{\Gamma, I} = \mathbb{P}(T_{\text{abe}} \geq \zeta_\alpha | \mathcal{Z})$ is the power of a one-sided α -level sensitivity analysis conducted with Γ of the aberrant rank test T_{abe} . Typically, $\Psi_{\Gamma, I}$ is computed with respect to draws from a data generating process in the favorable situation in which there is no hidden bias and there is a treatment effect.

Recall that the aberrant null H_0^A is a more appropriate null hypothesis to test than both \tilde{H}_0 and H_0 when there exists a designated cut-off for what constitutes an aberrant outcome but more severely aberrant outcomes are worse than less severely aberrant outcomes. Note that both the aberrant rank test and the Mantel-Haenszel test are valid for testing the aberrant null in the sense that they both have pivotal distributions if the aberrant null holds. When testing the aberrant null, among these two candidate tests, intuitively the aberrant rank test should be more appealing and natural to choose since its null distribution also incorporates the fact that the severity of aberration is fixed under the aberrant null, while the null distribution of the Mantel-Haenszel test under the aberrant null is the same as that un-

der the weaker null \tilde{H}_0 which only looks at the dichotomized outcome indicating aberration/non-aberration. However, as we will show in Sections 2.4.2 and 2.4.3, when testing the aberrant null in observational studies, although the aberrant rank test is indeed more powerful than the Mantel-Haenszel test in many cases, there also exist settings under which the Mantel-Haenszel test is instead more powerful. This motivates us to develop a new adaptive testing procedure in Section 2.5.2 and use it to combine the aberrant rank test and the Mantel-Haenszel test to guarantee that the resulting adaptive approach for testing the aberrant null uniformly dominates the Mantel-Haenszel test in large samples and performs well in finite samples.

2.4.2 Design sensitivity formula of the aberrant rank test

There is an extensive literature on deriving design sensitivity formulas for various test statistics in matched observational studies. These design sensitivity formulas provide powerful tools for asymptotically evaluating the performances of various tests in a sensitivity analysis. However, all previous approaches either require the use of pair matching (Rosenbaum, 2010, 2011; Hansen et al., 2014; Rosenbaum and Small, 2017; Howard and Pimentel, 2019; Fogarty et al., 2021), or require a particular structure for the test statistic to which many rank statistics do not conform (Rosenbaum, 2013, 2014). There are no design sensitivity formulas for popular test statistics, such as the Wilcoxon rank sum test and the Hodges-Lehmann aligned rank test, and the aberrant rank test discussed above; more generally, currently methods cannot handle test statistics where there are matched strata with multiple controls and ranking is done across matched strata as this induces dependence between matched strata that are typically assumed to be independent in many sensitivity analyses. In this section, we derive a novel design sensitivity formula for

the aberrant rank test, of which the Wilcoxon rank sum test is a special case. Our proof technique involves applying empirical processes to matched data. To the best of our knowledge, this is the first application of such machinery to the sensitivity analysis of matched observational studies and can potentially be used to study many other rank tests in the matched setting.

We need a few regularity and causal assumptions of responses \mathbf{R} under the alternative for deriving the design sensitivity. Without loss of generality, in Section 2.4.2 we suppose that in each stratum i , unit $j = 1$ received treatment and $j = 2, \dots, m$ received control. If not, we just need to simply reassign index $j = 1$ to the treated in each stratum.

Assumption 1 (i.i.d. strata). *The responses from each stratum i - (R_{i1}, \dots, R_{im}) are i.i.d. realizations from a continuous multivariate distribution $F(x_1, \dots, x_m)$.*

Assumption 1 implies existence of a super-population model for the potential outcomes, which is common in deriving design sensitivity values (Rosenbaum, 2004, 2010). We remark that this assumption as well as others below are only used to derive the design sensitivity formula; they are not required for the validity of the aberrant rank test defined in Section 2.4.1 or the adaptive approach introduced in Section 2.5.2.

Suppose that $F(x_1, \dots, x_m)$ has marginal cumulative distributions $F_1(x_1), \dots, F_m(x_m)$ and densities $f_1(x_1), \dots, f_m(x_m)$. Let $F_{(1)}, \dots, F_{(m)}$ be the associated marginal distribution with densities $f_{(1)}, \dots, f_{(m)}$ of $R_{i(1)} \leq \dots \leq R_{i(m)}$, the ordered responses within stratum i . The next two assumptions place regularity conditions on the marginal distributions of f_j and its ordered counterpart $f_{(j)}$.

Assumption 2 (Connectedness of the support). *For $j = 1, \dots, m$, let $s_j = \sup\{t :$*

$\mathbb{P}(R_{ij} \geq t) > 0\}$ (s_j can be ∞). Then $s_j > c$ and $f_j(t) > 0$ for any $t \in [c, s_j)$.

Assumption 3 ('Positive' and 'non-extreme' treatment effect). Let $s = \max_j s_j$. Then for any $t \in [c, s)$, we have $\mathbb{P}(R_{i1} \geq t) \geq \frac{1}{m} \sum_{j=1}^m \mathbb{P}(R_{ij} \geq t)$, and there exists an open interval $\mathcal{I} \subset [c, s)$ such that strict inequality holds for any $t \in \mathcal{I}$. Moreover, $\mathbb{P}(R_{i(m)} > R_{i1} \geq t) > 0$ for some $t \geq c$.

In words, Assumption 2 states that aberrant responses can be observed with non-zero probability, and the support of the distribution function of each individual's aberrant response is a connected set. Assumption 3 states that the aberrant response of the treated is stochastically larger than the average of the distribution function of all the responses within the same stratum, i.e., $\mathbb{P}(R_{i1} \geq t) \geq \frac{1}{m} \sum_{j=1}^m \mathbb{P}(R_{ij} \geq t)$. The remaining part of Assumption 3 is only intended to prevent design sensitivities from equaling 1 or going to ∞ ; see the proof of Theorem 1 in the Appendix A for details.

We give some examples to show that Assumptions 1-3 hold for many widely considered treatment effect models. In Examples 1-3 listed below, we assume that (R_{i1}, \dots, R_{im}) are i.i.d. continuous random vectors, and we assume that for each i , R_{i2}, \dots, R_{im} are identically distributed with the support equalling \mathbb{R} , and correlation of R_{ij_1} and R_{ij_2} is neither 1 or -1 for any two distinct $j_1, j_2 \in \{1, \dots, m\}$. ' \sim ' means two distributions are equal. We consider the following three examples: (i) **Example 1** (Additive treatment effects): $R_{i1} \sim R_{i2} + \beta$ for some $\beta > 0$ with $c \in \mathbb{R}$; (ii) **Example 2** (Multiplicative treatment effects): $R_{i1} \sim \delta \cdot R_{i2}$ for some $\delta > 1$ with $c > 0$; (iii) **Example 3** (Lehmann's alternative): $F_1 = p \cdot F_2^q + (1 - p) \cdot F_2$ for some $0 < p < 1$ and $q > 1$ with $c \in \mathbb{R}$, where $R_{i1} \sim F_1$ and $R_{i2} \sim F_2$. Lehmann's alternative is often used to model some uncommon but dramatic responses to treatment; see [Rosenbaum \(2010\)](#) (Chapter 16) for some real data examples.

Proposition 1. *Assumptions 1-3 hold for Examples 1-3.*

Theorem 1 (Design sensitivity of the aberrant rank test). *Define $G(v) = \frac{1}{m} \sum_{j=1}^m \max\{F_j(v) - F_j(c), 0\}$ and $[m-1] = \{1, \dots, m-1\}$. Under Assumptions 1-3,*

$$\mathbb{E} \left\{ \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \right\} = \mathbb{E}\{G(R_{i1})\} \quad (2.5)$$

has a unique solution for $\Gamma \in (1, +\infty)$, call it $\tilde{\Gamma}$. Then $\tilde{\Gamma}$ is the design sensitivity of the aberrant rank test as in (2.4). That is, as $I \rightarrow \infty$, the power $\Psi_{\Gamma, I}$ of a one-sided α -level sensitivity analysis satisfies $\Psi_{\Gamma, I} \rightarrow 1$ if $\Gamma < \tilde{\Gamma}$, and $\Psi_{\Gamma, I} \rightarrow 0$ if $\Gamma > \tilde{\Gamma}$.

Proofs of all propositions and theorems in this paper are provided in Appendix A in the supplementary materials. Theorem 1 confirms that the design sensitivity of the aberrant rank test depends only on the underlying data generating distribution F and is independent of the level α and the sample size I . Setting $c = -\infty$ gives the design sensitivity formula of the Wilcoxon rank sum test.

2.4.3 Asymptotic comparison via the design sensitivity

Theorem 1 allows us to numerically calculate the design sensitivity of the aberrant rank test in each situation, and compare it with that of the Mantel-Haenszel test. Since the design sensitivity only depends on the data generating process and is independent of the level α and the sample size I , it gives us an intrinsic and elegant measurement of how robust a test is to hidden bias, and enables us to asymptotically compare two tests for observational studies. For convenience, as in Theorem 1, in Section 2.4.3 we still assume that for each i , without loss of generality, $j = 1$ receives treatment and others receive control, and $(R_{i1}, \dots, R_{im})^T = (r_{Ti1}, r_{Ci2}, \dots, r_{Cim})^T$ is an i.i.d. realization from a multivariate

continuous distribution. To make our calculation easier and clearer, we further assume that: First, $r_{Tij} = g(r_{Cij})$ for some deterministic function g . That is, given g , r_{Tij} is only determined by r_{Cij} and is independent of other individuals' outcomes given r_{Cij} ; Second, each r_{Cij} is realized from the same distribution F ; Third, within the same stratum, R_{i1}, \dots, R_{im} are independent of each other. In Appendix B, we also examine the cases when R_{i1}, \dots, R_{im} are correlated. Note that these three assumptions merely serve to simplify the simulations and are not necessary for Theorem 1.

We consider the following four models: (i) **Model 1** (additive treatment effects, normal distribution): $r_{Tij} = r_{Cij} + \beta$, F is the standard normal distribution; (ii) **Model 2** (additive treatment effects, Laplace distribution): $r_{Tij} = r_{Cij} + \beta$, F is the Laplace distribution with mean zero and variance one; (iii) **Model 3** (multiplicative treatment effects, normal distribution): $r_{Tij} = \delta \cdot r_{Cij}$, F is the standard normal distribution; (iv) **Model 4** (multiplicative treatment effects, Laplace distribution): $r_{Tij} = \delta \cdot r_{Cij}$, F is the Laplace distribution with mean zero and variance one. For all four models, we set the aberrant response threshold to be $c = 1$, that is, any response $R_{ij} > 1$ is considered to be an aberrant response. Table 2.1 reports the design sensitivities of the Mantel-Haenszel test and the aberrant rank test under Models 1-4 with $m = 4$ (i.e., matching with three controls) and various β and δ . Calculation is based on Monte-Carlo simulations. Specifically, under each data generating model, we can calculate the left-hand side (LHS) of (2.5) for each fixed Γ and the right-hand side (RHS) of (2.5) using Monte-Carlo simulations. According to Lemma 9 in Appendix A, the RHS of (2.5) is a strictly monotonically increasing function of Γ , therefore we can use the bisection method to find the solution of equation (2.5). According to Theorem 1, that solution is exactly the design sensi-

tivity of the aberrant rank test given each data generating model.

Two clear patterns emerge in Table 2.1. First, the choice of the test statistic has a huge influence on the design sensitivities. For example, under Model 3 with $\delta = 2$, the design sensitivity of the aberrant rank test is nearly twice as big as that of the Mantel-Haenszel test. Second, whether or not the aberrant rank test outperforms the Mantel-Haenszel test depends upon the unknown data generating distribution of \mathcal{F} . As seen from Table 2.1, under Models 1, 3 and 4, the aberrant rank test should be asymptotically less sensitive to unmeasured confounders with larger design sensitivities; instead under Model 2, the Mantel-Haenszel test should be more favorable in a sensitivity analysis with larger $\tilde{\Gamma}$. These theoretical insights are validated in a simulation study in Section 2.6.

Table 2.1: Design sensitivities of the Mantel-Haenszel test and the aberrant rank test under Models 1-4 and matching with three controls with various parameters. The larger of the two design sensitivities of the two tests is in bold in each case.

| Test statistic | Model 1: additive, normal | | | Model 2: additive, Laplace | | |
|----------------|---------------------------------|-----------------|-----------------|----------------------------------|-----------------|-----------------|
| | $\beta = 0.50$ | $\beta = 0.75$ | $\beta = 1.00$ | $\beta = 0.50$ | $\beta = 0.75$ | $\beta = 1.00$ |
| M-H test | 2.36 | 3.56 | 5.30 | 2.36 | 3.91 | 7.21 |
| Aberrant rank | 2.63 | 4.20 | 6.50 | 2.28 | 3.59 | 5.93 |
| Test statistic | Model 3: multiplicative, normal | | | Model 4: multiplicative, Laplace | | |
| | $\delta = 1.50$ | $\delta = 1.75$ | $\delta = 2.00$ | $\delta = 1.50$ | $\delta = 1.75$ | $\delta = 2.00$ |
| M-H test | 1.80 | 2.11 | 2.37 | 1.75 | 2.07 | 2.37 |
| Aberrant rank | 2.50 | 3.28 | 4.07 | 2.15 | 2.75 | 3.36 |

We give some intuition as to why the aberrant rank test should sometimes be preferred over the Mantel-Haenszel test and other times the Mantel-Haenszel

should be preferred. Suppose that $r_{Cij} \stackrel{iid}{\sim} f_0(x)$ and $r_{Tij} \stackrel{iid}{\sim} f_1(x)$ where f_0 and f_1 are two densities. Roughly speaking, the more $f_1(x)/f_0(x)$ departs from 1, the easier it is to distinguish the treated and control given the outcome value x . For Model 1 with $\beta > 0$, $f_1(x)/f_0(x) = \exp(\beta x - \beta^2/2)$. For Model 2 with $\beta > 0$ and $x > \beta$, $f_1(x)/f_0(x) = \exp(\sqrt{2}\beta)$. For Model 3 with $\delta > 1$ and $x > 0$, $f_1(x)/f_0(x) = \delta^{-1} \exp((1 - \delta^{-2})x^2/2)$. For Model 4 with $\delta > 1$, $f_1(x)/f_0(x) = \delta^{-1} \exp(\sqrt{2}(1 - \delta^{-1})x)$. Thus, for Models 1, 3 and 4 with $\beta > 0$ and $\delta > 1$, suppose that c is large enough, specially $c \geq \max\{\frac{\beta}{2}, \sqrt{\frac{2 \log \delta}{1 - 1/\delta^2}}, \frac{\log \delta}{\sqrt{2}(1 - 1/\delta)}\}$, then $f_1(x)/f_0(x) \geq 1$ and $f_1(x)/f_0(x)$ is increasing for all $x \geq c$. That is, in these three models, it is easier to detect the true treatment effect at the tail (i.e., larger outcome value x) and the aberrant rank test should outperform the Mantel-Haenszel test by assigning larger weights to more aberrant responses (i.e., larger outcome values) via aberrant ranks. For Model 2 with $c \geq \beta$, $f_1(x)/f_0(x)$ is a constant for $x \geq c$. In this case, the Mantel-Haenszel test should be more powerful than the aberrant rank test since it does not distinguish different magnitudes of severity, while the aberrant rank test loses power by unnecessarily assigning unequal weights based upon the degree of aberration.

2.5 A New, General Adaptive Approach to Combine Two Test Statistics in Observational Studies

2.5.1 Motivation and previous methods

From the perspectives of design sensitivity and power of sensitivity analysis, neither the Mantel-Haenszel test nor the aberrant rank test uniformly dominates the other. Instead, which test is to be preferred depends upon the data generating

process. Unfortunately we typically do not know which one is better for a given setting since we do not typically know the true data generating process. This type of problem is common in observational studies, where we typically have several available tests that we can use, but there is no single choice that can dominate all other choices in all possible situations (Rosenbaum, 2012). To overcome this type of problem in observational studies, various methods have been proposed. Among these, for example, Heller et al. (2009) and Zhang et al. (2011) used a sample splitting method in which a fraction of the data, the planning sample, is used to select a test and the remaining part of the data, the analysis sample, to carry out a test. The sample splitting method throws out the planning sample for carrying out the test which reduces power for small or moderate sample sizes.

Rosenbaum (2012) proposed a data-driven, adaptive approach to combine two test statistics in matched observational studies. It does not require dropping samples for design, and is designed to achieve the larger of the two design sensitivities of the component tests with a smaller cost for multiplicity adjustment compared with the Bonferroni adjustment. However, this adaptive approach can only be applied to test statistics that are uniformly bounded by a known distribution, which typically requires either the matching to be by pairs or the outcomes to be binary, neither of which would hold for many commonly used tests; see Appendix H and Rosenbaum (2012) for details. For example, the existing approach cannot be used for the aberrant rank test, the Wilcoxon rank sum test, the Hodges-Lehmann aligned rank test or the Huber-Maritz m -tests (Gastwirth et al., 2000; Rosenbaum, 2002b, 2007).

2.5.2 A new, general adaptive test via two-stage programming

Instead of focusing on matching with $m - 1$ ($m \geq 2$) controls as we did in previous sections, in this section we consider a more general matching regime allowing matching with different number of controls across the strata. Suppose that there are I matched strata with n_i individuals in the i -th stratum, $N = \sum_{i=1}^I n_i$ individuals in total. $n_i = 2$ with $Z_{i1} + Z_{i2} = 1$ for all i refers to pair matching. $n_i = m \geq 3$ with $\sum_{j=1}^m Z_{ij} = 1$ for all i refers to matching with multiple controls. In full matching, n_i can take different values with different i , and $\sum_{j=1}^{n_i} Z_{ij} \in \{1, n_i - 1\}$ for all i (i.e., either one treated individual and one or more controls, or one control and one or more treated individuals, within each stratum). As in previous sections, we still let $\mathbf{Z} = (Z_{11}, \dots, Z_{In_I})^T$ be the binary vector of treatment assignments, and $\mathbf{Z} \in \mathcal{Z}$ if and only if $\sum_{j=1}^{n_i} Z_{ij} = 1$ for each i . The constraint $\sum_{j=1}^{n_i} Z_{ij} = 1$ for all i is no more restrictive than assuming $\sum_{j=1}^{n_i} Z_{ij} \in \{1, n_i - 1\}$ for all i and is only imposed to make our derivations in this section clearer. See Appendix E for the detailed description of how the procedure derived in this section can be directly extended to allow for $\sum_{j=1}^{n_i} Z_{ij} \in \{1, n_i - 1\}$ for all i . We still let \mathcal{F} be the set of all fixed quantities of r_{Tij} , r_{Cij} , \mathbf{x}_{ij} and u_{ij} .

Motivated by the demand of performing adaptive inference in much more general settings than the traditional adaptive approach, we develop here a new adaptive approach that can combine any two sum test statistics which refer to any test statistics with the form $T = \mathbf{Z}^T \mathbf{q} = \sum_{i=1}^I \sum_{j=1}^{n_i} Z_{ij} q_{ij}$ where each q_{ij} is an arbitrary function of the response vector $\mathbf{R} = (R_{11}, \dots, R_{In_I})^T$ that does not vary with $\mathbf{Z} \in \mathcal{Z}$ under the null hypothesis, and can work under various matching strategies due to the flexibility of the value of each n_i . We would like the power of the adaptive test to be asymptotically no less than the higher of the two powers of the component tests

in sensitivity analysis. The idea is that when the sample size is large, to achieve the higher of the two powers of the component tests is almost equivalent to achieving the larger of the two design sensitivities of the component tests. Consider applying the Bonferroni adjustment to the component tests $T_k = \mathbf{Z}^T \mathbf{q}_k = \sum_{i=1}^I \sum_{j=1}^{n_i} Z_{ij} q_{ijk}$ with $\mathbf{q}_k = (q_{11k}, \dots, q_{In_1k})^T$ where each q_{ijk} is a function of the response vector \mathbf{R} and $k \in \{1, 2\}$. Let $p_{ij} = \mathbb{P}(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = \exp(\gamma u_{ij}) / \sum_{j'=1}^{n_i} \exp(\gamma u_{ij'})$. For a one-sided test with level α and given Γ ,

$$\text{the Bonferroni adjustment rejects the null if } \max_{k \in \{1, 2\}} \min_{\mathbf{u} \in \mathcal{U}} \frac{t_k - \mu_{k, \mathbf{u}}}{\sigma_{k, \mathbf{u}}} \geq \Phi^{-1}(1 - \alpha/2), \quad (2.6)$$

where t_k is the observed value of T_k , and $\mu_{k, \mathbf{u}} = \mathbb{E}_{\Gamma, \mathbf{u}}(\mathbf{Z}^T \mathbf{q}_k \mid \mathcal{F}, \mathcal{Z}) = \sum_{i=1}^I \sum_{j=1}^{n_i} p_{ij} q_{ijk}$ and $\sigma_{k, \mathbf{u}}^2 = \text{Var}_{\Gamma, \mathbf{u}}(\mathbf{Z}^T \mathbf{q}_k \mid \mathcal{F}, \mathcal{Z}) = \sum_{i=1}^I \sum_{j=1}^{n_i} p_{ij} q_{ijk}^2 - \sum_{i=1}^I (\sum_{j=1}^{n_i} p_{ij} q_{ijk})^2$ are the expectations and variances of T_k under the null hypothesis, with a specified Γ and given all unobserved covariates $\mathbf{u} = (u_{11}, \dots, u_{In_1})^T \in [0, 1]^N =: \mathcal{U}$. The term $\alpha/2$ in the RHS of (2.6) comes from the Bonferroni adjustment with two component tests. Under a normal approximation, the standard deviate of t_k follows a standard normal distribution, thus (2.6) is a valid testing procedure with level α and given Γ in a sensitivity analysis. Note that the design sensitivity of a test only depends on the data generating distribution and is independent of level α . Using an argument parallel to the proof of Proposition 2 in Rosenbaum (2012), it is straightforward to show that applying (2.6) with the two component tests can achieve the larger of the two design sensitivities, where we reject the null as long as one of the two tests rejects the null with significant level $\alpha/2$. However, simply applying (2.6) may lose power due to two significant deficiencies. First, it does not use the fact that the confounder has to impact the treatment assignment in the same way between the two component tests on the same

outcome variable. Second, it does not incorporate the information of the correlation between the two component tests. We implement a two-stage programming procedure to overcome these two deficiencies.

In the first stage, we utilize bounds on the correlation between T_1 and T_2 to replace $\Phi^{-1}(1 - \alpha/2)$ with a smaller rejection threshold under the given Γ and level $\alpha < 1/2$. Under some mild regularity conditions (see Appendix C for details), (T_1, T_2) is asymptotically bivariate normal in the sense that for large I , the distribution function of $\left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}}\right)$ can be approximated by that of $(X_1, X_2) \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{\mathbf{u}} \\ \rho_{\mathbf{u}} & 1 \end{pmatrix}\right)$, where $\rho_{\mathbf{u}} = \mathbb{E}\left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}} \cdot \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \middle| \mathcal{F}, \mathcal{Z}\right)$ can be expressed as

$$\rho_{\mathbf{u}} = \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} p_{ij} q_{ij1} q_{ij2} - \sum_{i=1}^I (\sum_{j=1}^{n_i} p_{ij} q_{ij1}) (\sum_{j=1}^{n_i} p_{ij} q_{ij2})}{\sqrt{\sum_{i=1}^I \sum_{j=1}^{n_i} p_{ij} q_{ij1}^2 - \sum_{i=1}^I (\sum_{j=1}^{n_i} p_{ij} q_{ij1})^2} \sqrt{\sum_{i=1}^I \sum_{j=1}^{n_i} p_{ij} q_{ij2}^2 - \sum_{i=1}^I (\sum_{j=1}^{n_i} p_{ij} q_{ij2})^2}}. \quad (2.7)$$

Let $Q_{\rho_{\mathbf{u}}, \alpha}$ be the quantile such that $\mathbb{P}(X_1 \leq Q_{\rho_{\mathbf{u}}, \alpha}, X_2 \leq Q_{\rho_{\mathbf{u}}, \alpha}) = 1 - \alpha$. Note that we would like to derive a valid testing procedure given any \mathbf{u} with the given Γ and α , we should look at the worst-case rejection threshold $\max_{\mathbf{u} \in \mathcal{U}} Q_{\rho_{\mathbf{u}}, \alpha}$. Invoking Slepian's lemma (Slepian, 1962), to find $\max_{\mathbf{u} \in \mathcal{U}} Q_{\rho_{\mathbf{u}}, \alpha}$, it suffices to find $\min_{\mathbf{u} \in \mathcal{U}} \rho_{\mathbf{u}}$. Through setting $w_{ij} = \exp(\gamma u_{ij})$, we further transform solving $\min_{\mathbf{u} \in \mathcal{U}} \rho_{\mathbf{u}}$ into solving

$$\begin{aligned} & \underset{w_{ij}}{\text{minimize}} \quad \rho_{\mathbf{u}} \quad (*) \\ & \text{subject to} \quad 1 \leq w_{ij} \leq \Gamma, \quad \forall i, j \end{aligned}$$

where $\rho_{\mathbf{u}}$ is as in (2.7) with $p_{ij} = w_{ij} / \sum_{j'=1}^{n_i} w_{ij'}$. (*) is a large-scale nonlinear optimization problem with box constraints which can be solved approximately in a reasonable amount of time by the well-known L-BFGS-B algorithm, which is a

limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) algorithm allowing box constraints (Byrd et al., 1995). Denote the optimal value of (*) with sensitivity parameter Γ as ρ_Γ^* . Then the corresponding worst-case quantile $\max_{\mathbf{u} \in \mathcal{U}} Q_{\rho_{\mathbf{u}}, \alpha}$ equals $Q_{\rho_\Gamma^*, \alpha}$ by Slepian's lemma. It is well known that $Q_{\rho_\Gamma^*, \alpha} < \Phi^{-1}(1 - \alpha/2)$ as long as $\rho_\Gamma^* > -1$. Thus, for two positively correlated test statistics T_1 and T_2 , especially when the correlation is much greater than zero (which is the case when combining the Mantel-Haenszel test and the aberrant rank test), $Q_{\rho_\Gamma^*, \alpha}$ is a much less conservative rejection threshold than $\Phi^{-1}(1 - \alpha/2)$.

In the second stage, we apply the minimax procedure developed in Fogarty and Small (2016) to replace the test statistic $\max_{k \in \{1,2\}} \min_{\mathbf{u} \in \mathcal{U}} (t_k - \mu_{k,\mathbf{u}}) / \sigma_{k,\mathbf{u}}$ in (2.6) with a larger one. Note that the following max-min inequality always holds

$$\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1,2\}} \frac{t_k - \mu_{k,\mathbf{u}}}{\sigma_{k,\mathbf{u}}} \geq \max_{k \in \{1,2\}} \min_{\mathbf{u} \in \mathcal{U}} \frac{t_k - \mu_{k,\mathbf{u}}}{\sigma_{k,\mathbf{u}}}, \quad (2.8)$$

and strict inequality is possible. (2.8) implies that instead of performing the two sensitivity analyses to solve $\min_{\mathbf{u} \in \mathcal{U}} (t_k - \mu_{k,\mathbf{u}}) / \sigma_{k,\mathbf{u}}$ for $k \in \{1,2\}$ separately, we should conduct a simultaneous sensitivity analysis to directly check if $\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1,2\}} (t_k - \mu_{k,\mathbf{u}}) / \sigma_{k,\mathbf{u}} \geq Q_{\rho_\Gamma^*, \alpha}$ - if the inequality holds, we reject the null; otherwise, we fail to reject. Adapting the one-sided minimax procedure described in Part B of the Appendices of Fogarty and Small (2016) with our new rejection threshold $Q_{\rho_\Gamma^*, \alpha}$, this procedure can be implemented through setting $s_i = 1 / \sum_{j'=1}^{n_i} \exp(\gamma u_{ij'})$ and solving the following quadratically constrained linear program with M being a sufficiently large constant (see Appendix D for the

detailed derivation):

$$\begin{aligned}
& \underset{y, p_{ij}, s_i, b_k}{\text{minimize}} && y \quad (**) \\
& \text{subject to} && y \geq (t_k - \mu_{k,\mathbf{u}})^2 - Q_{\rho_{\Gamma}^*, \alpha}^2 \sigma_{k,\mathbf{u}}^2 - Mb_k \quad \forall k \in \{0, 1\} \\
& && \sum_{j=1}^{n_i} p_{ij} = 1 \quad \forall i \\
& && s_i \leq p_{ij} \leq \Gamma s_i \quad \forall i, j \\
& && p_{ij} \geq 0 \quad \forall i, j \\
& && b_k \in \{0, 1\} \quad \forall k \in \{0, 1\} \\
& && -Mb_k \leq t_k - \mu_{k,\mathbf{u}} \leq M(1 - b_k), \quad \forall k \in \{0, 1\}
\end{aligned}$$

and checking whether the optimal value $y_{\Gamma}^* \geq 0$. If it is, we reject the null; otherwise, we fail to reject. The ‘ M ’ constraint here precludes a directional error, as without it one might reject the null if evidence pointed in the opposite direction of the alternative. A quadratically constrained linear program can be efficiently solved with many available solvers. Contrary to implementing (*), from which the gains in power is relatively large when the correlation between T_1 and T_2 is strong, implementing (**) (the minimax procedure) typically can have marked improvement of power when the correlation between T_1 and T_2 is weak; see Section 8 in [Fogarty and Small \(2016\)](#). That is, by implementing our two-step programming (*) and (**), we can always expect gains in power no matter the correlation between the two component tests are strong or weak.

To conclude, in our new adaptive test implemented via Algorithm 1 listed below,

$$\text{we reject the null if } \min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1, 2\}} \frac{t_k - \mu_{k,\mathbf{u}}}{\sigma_{k,\mathbf{u}}} \geq Q_{\rho_{\Gamma}^*, \alpha}. \quad (2.9)$$

Algorithm 1: Two-stage programming for implementing (2.9) as an adaptive test

Input: Sensitivity parameter Γ ; level α of the one-sided test; treatment

assignment indicators $\mathbf{Z} = (Z_{11}, \dots, Z_{In_1})^T$; the two score vectors

$\mathbf{q}_1 = (q_{111}, \dots, q_{In_11})^T$ and $\mathbf{q}_2 = (q_{112}, \dots, q_{In_12})^T$ associated with

$T_1 = \sum_{i=1}^I \sum_{j=1}^{n_i} Z_{ij} q_{ij1}$ and $T_2 = \sum_{i=1}^I \sum_{j=1}^{n_i} Z_{ij} q_{ij2}$ respectively ;

Step 1: Solve (*) to get the worst-case correlation ρ_Γ^* along with the

corresponding worst-case quantile $Q_{\rho_\Gamma^*, \alpha}$;

Step 2: Solve (**) with $Q_{\rho_\Gamma^*, \alpha}$ obtained from Step 1, and get the corresponding

optimal value y_Γ^* ;

Output: If $y_\Gamma^* \geq 0$, we reject the null; otherwise, we fail to reject.

When $\Gamma = 1$, the testing procedure (2.9) implemented via Algorithm 1 reduces to the usual testing procedure with the maximum statistic $\max\{T_1, T_2\}$ with correcting for $\text{Cor}(T_1, T_2)$. An R package SuperAdap for implementing the two-stage programming method described in Algorithm 1 is posted at <https://github.com/siyuheng/SuperAdap>. Proposition 2 says that the sensitivity analysis with the adaptive testing procedure described in Algorithm 1 has the correct level α asymptotically.

Proposition 2. For any unknown true $\mathbf{u}_0 \in \mathcal{U}$ and true $\Gamma_0 \leq \Gamma$, we have

$$\lim_{I \rightarrow \infty} \mathbb{P}_{\Gamma_0, \mathbf{u}_0} \left(\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1, 2\}} \frac{t_k - \mu_{k, \mathbf{u}}}{\sigma_{k, \mathbf{u}}} \geq Q_{\rho_\Gamma^*, \alpha} \middle| \mathcal{F}, \mathcal{Z} \right) \leq \alpha.$$

A nice feature of the traditional adaptive test (Rosenbaum, 2012) is that its design sensitivity is the larger of the two component tests. The Bonferroni adjustment and sample splitting method also have this design sensitivity but sometimes lose

power in finite samples to the adaptive test. In Theorem 2, we prove that the design sensitivity of our new adaptive approach is always greater than or equal to both two design sensitivities of the component tests, and surprisingly, *strict inequality is possible*. We refer to this new phenomenon as “super-adaptivity.”

Theorem 2 (Super-adaptivity). *Let $\tilde{\Gamma}_1$ and $\tilde{\Gamma}_2$ be the two design sensitivities of the two tests T_1 and T_2 , and let $\tilde{\Gamma}_{1:2}$ be the design sensitivity of the adaptive testing procedure (2.9) implemented by Algorithm 1 with T_1 and T_2 as the two component tests. We have $\tilde{\Gamma}_{1:2} \geq \max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$, and strict inequality is possible.*

Theorem 2 shows that in terms of the design sensitivity, our new adaptive test dominates all the existing methods, including the traditional adaptive test, the Bonferroni adjustment and sample splitting. Recall that the design sensitivity is a threshold of the consistency of a test in a sensitivity analysis with respect to sensitivity parameter Γ . When $\tilde{\Gamma}_{1:2} = \max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$, roughly speaking, the new adaptive test is consistent as long as one of the two component tests was consistent, which can also be obtained by the traditional adaptive approach. When $\tilde{\Gamma}_{1:2} > \max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$, the new adaptive test can still be consistent even if neither of the two component tests was consistent, which cannot be achieved from using the traditional adaptive approach.

Typically, substantial gains in design sensitivity (i.e., gaps between $\tilde{\Gamma}_{1:2}$ and $\max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$) resulting from Algorithm 1 are more likely to be observed with two negatively correlated, independent or weakly positively correlated component statistics than with two strongly positively correlated component statistics (see Table 2.5 in Appendix A). When combining two statistics T_1 and T_2 on one response vector \mathbf{R} in an adaptive test, we often expect T_1 and T_2 to be highly positively correlated, in which case gains in design sensitivity may be hard to see without large

samples. But Theorem 2 is still worth highlighting since it is the first time that an adaptive test can result in a design sensitivity strictly larger than $\max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$, and it can inspire further studies on designing new adaptive tests with larger design sensitivities.

Note that the design sensitivity only measures limiting insensitivity to hidden bias. In terms of the finite sample power, we need to pay the price for correcting for the two component tests in the adaptive test. That is, Theorem 2 does not imply that the power of the adaptive test in a sensitivity analysis is always greater than or equal to the maximal power of the two component tests for any sample size. Instead, Theorem 2 implies that as long as the sample size is sufficiently large, applying Algorithm 1 to perform an adaptive inference is as good or better than knowing which of the two component tests should be better and using only that test, and is typically much better than incorrectly choosing the worse one among the two component tests, regardless of what the unknown data generating process is and what the two component tests are. Having theoretically derived this favorable asymptotic property of the adaptive test in Theorem 2, we turn to examining its performance with realistic sample sizes via simulations in Section 2.6 and Tables 2.4 and 2.5 in Appendix A.

2.6 Simulation Studies

We examine the finite sample power of sensitivity analyses to check the validity of the theoretical intuitions gained from calculating design sensitivities and compare the performances of (i) the Mantel-Haenszel test, (ii) the aberrant rank test, and (iii) our new adaptive test applying Algorithm 1 with the Mantel-Haenszel test and the aberrant rank test as components. That is, we use simulations to estimate

the probability that the worst-case p-value given by a test statistic in a sensitivity analysis with sensitivity parameter Γ will be less than $\alpha = 0.05$ under the favorable situation when there is an actual treatment effect and no hidden bias. Table 2.2 summarizes the simulated power of the three tests under Models 1 - 4 discussed in Section 2.4.3, where we match with three controls and number of matched strata $I = 100$ or $I = 1000$. In Table 2.2, we set $\beta = 1$ for Models 1 and 2 and set $\delta = 2$ for Models 3 and 4. For reference, we also give the design sensitivity of each test statistic in the first row of each block. We summarize the simulated size of the above three tests in Table 2.7 in Appendix F.

Table 2.2: Simulated power of the Mantel-Haenszel test, the aberrant rank test and the adaptive test. We set $\alpha = 0.05$, $c = 1$ and $m = 4$. We set $\beta = 1$ for Models 1 and 2 and $\delta = 2$ for Models 3 and 4. Each number is based on 2,000 replications. The largest of the three simulated powers in each case is in bold.

| Model 1 | $I = 100$ Matched Strata | | | $I = 1000$ Matched Strata | | |
|------------------|--------------------------|-------------|-------------|---------------------------|-------------|-------------|
| | M-H test | Aberrant | Adaptive | M-H test | Aberrant | Adaptive |
| $\tilde{\Gamma}$ | 5.30 | 6.50 | ≥ 6.50 | 5.30 | 6.50 | ≥ 6.50 |
| $\Gamma = 3.0$ | 0.71 | 0.87 | 0.83 | 1.00 | 1.00 | 1.00 |
| $\Gamma = 3.5$ | 0.46 | 0.70 | 0.63 | 1.00 | 1.00 | 1.00 |
| $\Gamma = 4.0$ | 0.28 | 0.52 | 0.43 | 0.96 | 1.00 | 1.00 |
| $\Gamma = 4.5$ | 0.14 | 0.32 | 0.25 | 0.61 | 0.99 | 0.99 |
| $\Gamma = 5.0$ | 0.08 | 0.21 | 0.14 | 0.17 | 0.89 | 0.82 |
| $\Gamma = 5.5$ | 0.04 | 0.12 | 0.09 | 0.02 | 0.57 | 0.47 |
| $\Gamma = 6.0$ | 0.01 | 0.06 | 0.04 | 0.00 | 0.20 | 0.14 |
| Model 2 | $I = 100$ Matched Strata | | | $I = 1000$ Matched Strata | | |
| | M-H test | Aberrant | Adaptive | M-H test | Aberrant | Adaptive |
| $\tilde{\Gamma}$ | 7.21 | 5.93 | ≥ 7.21 | 7.21 | 5.93 | ≥ 7.21 |

| | | | | | | |
|----------------|-------------|------|------|-------------|------|------|
| $\Gamma = 3.0$ | 0.95 | 0.78 | 0.92 | 1.00 | 1.00 | 1.00 |
| $\Gamma = 3.5$ | 0.84 | 0.58 | 0.77 | 1.00 | 1.00 | 1.00 |
| $\Gamma = 4.0$ | 0.70 | 0.38 | 0.60 | 1.00 | 1.00 | 1.00 |
| $\Gamma = 4.5$ | 0.51 | 0.22 | 0.44 | 1.00 | 0.91 | 1.00 |
| $\Gamma = 5.0$ | 0.37 | 0.13 | 0.26 | 1.00 | 0.58 | 0.99 |
| $\Gamma = 5.5$ | 0.22 | 0.07 | 0.16 | 0.93 | 0.20 | 0.88 |
| $\Gamma = 6.0$ | 0.15 | 0.04 | 0.10 | 0.64 | 0.03 | 0.58 |

| Model 3 | <i>I</i> = 100 Matched Strata | | | <i>I</i> = 1000 Matched Strata | | |
|------------------|-------------------------------|-------------|-------------|--------------------------------|-------------|-------------|
| | M-H test | Aberrant | Adaptive | M-H test | Aberrant | Adaptive |
| $\tilde{\Gamma}$ | 2.37 | 4.07 | ≥ 4.07 | 2.37 | 4.07 | ≥ 4.07 |
| $\Gamma = 1.0$ | 0.94 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 |
| $\Gamma = 1.5$ | 0.52 | 0.94 | 0.93 | 1.00 | 1.00 | 1.00 |
| $\Gamma = 2.0$ | 0.15 | 0.74 | 0.71 | 0.63 | 1.00 | 1.00 |
| $\Gamma = 2.5$ | 0.04 | 0.47 | 0.39 | 0.01 | 1.00 | 1.00 |
| $\Gamma = 3.0$ | 0.01 | 0.24 | 0.17 | 0.00 | 0.94 | 0.89 |

| Model 4 | <i>I</i> = 100 Matched Strata | | | <i>I</i> = 1000 Matched Strata | | |
|------------------|-------------------------------|-------------|-------------|--------------------------------|-------------|-------------|
| | M-H test | Aberrant | Adaptive | M-H test | Aberrant | Adaptive |
| $\tilde{\Gamma}$ | 2.37 | 3.36 | ≥ 3.36 | 2.37 | 3.36 | ≥ 3.36 |
| $\Gamma = 1.0$ | 0.89 | 0.97 | 0.95 | 1.00 | 1.00 | 1.00 |
| $\Gamma = 1.5$ | 0.46 | 0.78 | 0.72 | 1.00 | 1.00 | 1.00 |
| $\Gamma = 2.0$ | 0.14 | 0.47 | 0.39 | 0.55 | 1.00 | 1.00 |
| $\Gamma = 2.5$ | 0.03 | 0.22 | 0.17 | 0.02 | 0.88 | 0.83 |
| $\Gamma = 3.0$ | 0.01 | 0.08 | 0.05 | 0.00 | 0.29 | 0.24 |

In Table 2.2, in general, the power increases as the number of matched strata I increases, and the power decreases as the bias magnitude Γ increases, which agrees with empirical knowledge. The simulated power also verifies the validity of our design sensitivity formula. That is, as $I \rightarrow \infty$, the power of the test in a sensitivity analysis goes to 1 for $\Gamma < \tilde{\Gamma}$, and the power goes to 0 for $\Gamma > \tilde{\Gamma}$. For example, see the row $\Gamma = 5.5$ for Model 1 in Table 2.2, as I increases from 100 to 1000, the power of the aberrant rank test with $\tilde{\Gamma} = 6.5 > 5.5$ is closer to 1, but the power of the Mantel-Haenszel test with $\tilde{\Gamma} = 5.3 < 5.5$ is closer to 0. From Table 2.2, we can also observe that in Models 1, 3 and 4, the aberrant rank test is more powerful than the Mantel-Haenszel test; instead, in Model 2 the Mantel-Haenszel test has higher power than the aberrant rank test, and the gap between the two powers of these two tests could be extremely large, especially with large sample size and sensitivity parameter Γ considerably greater than 1. For example, see Models 1 and 2 with $\Gamma = 5.5$ and $I = 1000$, and Models 3 and 4 with $\Gamma = 2.5$ and $I = 1000$. This confirms the two key insights obtained from calculation of design sensitivities: power of a sensitivity analysis can differ a lot with different choices between the two tests and the optimal choice between the two tests could be different under different data generating processes.

We now examine the asymptotic property of the adaptive test. Let $\tilde{\Gamma}_1$, and $\tilde{\Gamma}_2$ denote the design sensitivities of the Mantel-Haenszel test and the aberrant rank test respectively. As long as the given $\Gamma < \max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$, the power of the adaptive test in a sensitivity analysis goes to 1 as sample size $I \rightarrow \infty$, even if one of the powers of the two component tests goes to zero if $\min\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\} < \Gamma < \max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$. For example, see the rows $\Gamma = 6.0$ of Models 1 and 2 and the rows $\Gamma = 3.0$ of Models 3 and 4 in Table 2.2.

We then examine the finite sample performance of the adaptive test. As discussed in Section 2.5.2, although the adaptive test uniformly dominates the two component tests in terms of the design sensitivity, in practice we need to pay the price for correcting for the two component tests and the finite sample power of the adaptive test may sit in between those of the two component tests, which is the case in Table 2.2. If this is the case, the simulation results in Table 2.2 confirm that the price paid for correcting for the two component tests is very much worth it in the sense that if the power of the Mantel-Haenszel test and that of the aberrant rank test differ a lot, then the power of the adaptive test is typically much closer to the higher one of the powers of the two component tests than to the lower one in each case. This favorable finite sample property of the adaptive test holds both for relatively large sample sizes (e.g., see the cases $\Gamma = 5, I = 1000$ in Models 1 and 2) and relatively small sample sizes (e.g., see the cases $\Gamma = 1.5, I = 100$ in Models 3 and 4). To conclude, the new adaptive test is like a high quality insurance policy: we will lose a little money (the low cost of the insurance) if we bought one but an accident never occurs (i.e., if we were lucky enough to always choose the better one among the two component tests), but we will lose much more if an accident indeed occurs (i.e., if we unfortunately choose the worse one among the two component tests) but we never bought one.

2.7 Adaptive Inference of the Effect of Mother's Age on Child Stunted Growth

For the study of the effect of mother's age on child stunting discussed in Section 2.1.1.1, we summarize the worst-case p-values of a sensitivity analysis reported by three different test statistics: the Mantel-Haenszel test, the aberrant rank

test and the adaptive test applying Algorithm 1 putting together these two tests with various sensitivity parameters Γ ranging from 1.00 to 1.45; see Appendix G for more details. From Table 2.3, we find that the Mantel-Haenszel test fails to detect a possible treatment effect (i.e., worst-case p-value > 0.05) with sensitivity parameter $\Gamma = 1.17$ under level 0.05. However, the aberrant rank test can detect a possible treatment effect (i.e., worst-case p-value < 0.05) up to a much larger sensitivity parameter $\Gamma = 1.43$. Thus, we can see that when studying causal determinants of aberrant response, the aberrant rank test might be preferred to the Mantel-Haenszel test since it might be less sensitive. However, we did not know this in advance of looking at the data, and choosing the test that is less sensitive on the data will inflate Type I errors in a sensitivity analysis. To use the data in choosing the best test while controlling the Type I error rate, we apply the adaptive approach developed in Section 2.5.2 to combine the aberrant rank test with the Mantel-Haenszel test to guarantee a powerful test in sensitivity analyses. From Table 2.3, we can find that if we combine these two tests with the new adaptive approach, we can successfully detect the possible actual treatment effect with $\Gamma = 1.36$, which is close to the results obtained by using the more favorable one between the two component tests - the aberrant rank test, and substantially better than the least favorable of the two tests. Therefore, both the aberrant rank test and the adaptive test enable us to detect a significant treatment effect even with a nontrivial magnitude of hidden bias. Meanwhile, for this particular data set, the Mantel-Haenszel test would possibly give an exaggerated report of sensitivity to bias. This agrees with all our theoretical insights and simulations results.

Table 2.3: One-sided worst-case p-values under various Γ . The p-values ≈ 0.05 are in bold. We also report the approximate sensitivity value of each test with level 0.05.

| One-sided worst-case p-values under various Γ | | | |
|--|-----------------|---------------|---------------|
| | Mantel-Haenszel | Aberrant rank | Adaptive test |
| $\Gamma = 1.00$ | 0.010 | 0.001 | 0.001 |
| $\Gamma = 1.05$ | 0.017 | 0.001 | 0.003 |
| $\Gamma = 1.10$ | 0.028 | 0.003 | 0.005 |
| $\Gamma = 1.15$ | 0.043 | 0.005 | 0.008 |
| $\Gamma = 1.17$ | 0.051 | 0.006 | 0.010 |
| $\Gamma = 1.20$ | 0.064 | 0.008 | 0.014 |
| $\Gamma = 1.25$ | 0.089 | 0.013 | 0.022 |
| $\Gamma = 1.30$ | 0.121 | 0.020 | 0.032 |
| $\Gamma = 1.35$ | 0.157 | 0.029 | 0.047 |
| $\Gamma = 1.36$ | 0.165 | 0.031 | 0.050 |
| $\Gamma = 1.40$ | 0.198 | 0.040 | 0.064 |
| $\Gamma = 1.43$ | 0.225 | 0.049 | 0.077 |
| $\Gamma = 1.45$ | 0.244 | 0.055 | 0.086 |
| Sensitivity value | 1.17 | 1.43 | 1.36 |

2.8 Discussion

We have developed an adaptive aberrant rank approach to conducting inference about the effect of a treatment on aberrant (bad) outcomes from matched observational studies when there is an established cutoff for what constitutes an aberrant outcome but more aberrant outcomes are worse than less aberrant ones. We have

shown that our new approach asymptotically dominates the traditional approach (performing the Mantel-Haenszel test with the dichotomous outcome indicating aberration/non-aberration) and performs well in simulation studies. To establish the new approach, we have developed an empirical process approach to studying design sensitivity and developed a general adaptive testing procedure. These developments can be applied to other types of general matched observational studies beyond the aberrant outcome setting we have studied.

There are limitations to this work. For example, we have not discussed how to enable adjustment for measured variables that were not used for matching. This side information, along with the matched observed covariates, can potentially be used to perform covariance adjustment in randomization inference ([Rosenbaum, 2002a](#)). However, existing model-based covariance adjustment approaches, e.g., covariance adjustment with robust linear regression considered in [Rosenbaum \(2002a\)](#), may not be directly applicable in our setting since the aberrant rank considers some truncated outcome (zero if not aberrant and multi-valued if aberrant). It might be fruitful for future research to explore how to incorporate both matched and unmatched measured variables to perform some covariance adjustment to further develop the aberrant rank approach.

2.9 Appendices

Appendix A: Proofs and Related Simulations

Proof of Proposition 1

Proof. The validity of Assumptions 1 and 2 for each example follows immediately from the general assumptions on (R_{i1}, \dots, R_{im}) , so we just need to check the validity of Assumption 3. For Example 1 with $\beta > 0$ and $c \in \mathbb{R}$, $\mathbb{P}(R_{i1} \geq t) = \mathbb{P}(R_{i2} \geq t - \beta) > \mathbb{P}(R_{i2} \geq t)$ for all $t \geq c$. For Example 2 with $\delta > 1$ and $c > 0$, $\mathbb{P}(R_{i1} \geq t) = \mathbb{P}(R_{i2} \geq \frac{t}{\delta}) > \mathbb{P}(R_{i2} \geq t)$ for all $t \geq c$. For Example 3 with $0 < p < 1$, $q > 1$ and $c \in \mathbb{R}$, $\mathbb{P}(R_{i1} \geq t) = 1 - F_1(t) = 1 - p \cdot F_2^q(t) - (1 - p) \cdot F_2(t) > 1 - F_2(t) = \mathbb{P}(R_{i2} \geq t)$ for all $t \geq c$. Thus, in Examples 1-3, $\mathbb{P}(R_{i1} \geq t) > \frac{1}{m} \cdot \{\mathbb{P}(R_{i1} \geq t) + (m - 1) \cdot \mathbb{P}(R_{i2} \geq t)\} = \frac{1}{m} \sum_{j=1}^m \mathbb{P}(R_{ij} \geq t)$ holds true for any $t \geq c$. From the general assumptions on (R_{i1}, \dots, R_{im}) , $\mathbb{P}(R_{i(m)} > R_{i1} \geq t)$ for some $t \geq c$ is trivially true. Thus, Assumption 3 also holds for Examples 1-3. \square

Proof of Theorem 1

Lemma 1. Let $G(v) = \frac{1}{m} \sum_{j=1}^m \max\{F_j(v) - F_j(c), 0\}$. Under Assumption 1, we have as $I \rightarrow \infty$,

$$\sup_v \left| \frac{q(v | \mathbf{R})}{mI} - G(v) \right| \xrightarrow{a.s.} 0.$$

Proof. We have the following expression

$$\begin{aligned} & \sup_v \left| \frac{q(v | \mathbf{R})}{mI} - G(v) \right| \\ &= \sup_v \left| \frac{1}{m} \sum_{j'=1}^m \left[\frac{1}{I} \sum_{i'=1}^I \mathbb{1}(v \geq R_{i'j'} > c) - \max\{F_{j'}(v) - F_{j'}(c), 0\} \right] \right| \\ &\leq \sup_v \frac{1}{m} \sum_{j'=1}^m \left| \frac{1}{I} \sum_{i'=1}^I \mathbb{1}(v \geq R_{i'j'} > c) - \max\{F_{j'}(v) - F_{j'}(c), 0\} \right| \\ &\leq \frac{1}{m} \sum_{j'=1}^m \sup_v \left| \frac{1}{I} \sum_{i'=1}^I \mathbb{1}(v \geq R_{i'j'} > c) - \max\{F_{j'}(v) - F_{j'}(c), 0\} \right|. \end{aligned}$$

First, for each j' , we have

$$\mathbb{E}\{\mathbb{1}(v \geq R_{i'j'} > c)\} = \mathbb{P}(v \geq R_{i'j'} > c) = \max\{F_{j'}(v) - F_{j'}(c), 0\}.$$

Second, for each j' in the above sum, the bracketing number of $\mathbb{1}(v \geq R_{i'j'} > c)$ is bounded by Example 19.6 in Van der Vaart (2000) where we replace $t_0 = -\infty$ with $t_0 = c$. Combining these two facts together, for each j' , each \sup_v term goes to zero a.s. and we have the desired result. \square

Lemma 2. *Under Assumption 1, we have as $I \rightarrow \infty$,*

$$\frac{q(R_{ij} | \mathbf{R})}{mI} \xrightarrow{a.s.} G(R_{ij}) \quad \text{and} \quad \frac{q(R_{i(j)} | \mathbf{R})}{mI} \xrightarrow{a.s.} G(R_{i(j)}).$$

Proof. The conclusion follows immediately from Lemma 1 and the fact that $|\frac{q(R_{ij} | \mathbf{R})}{mI} - G(R_{ij})| \leq \sup_v \left| \frac{q(v | \mathbf{R})}{mI} - G(v) \right|$ and $|\frac{q(R_{i(j)} | \mathbf{R})}{mI} - G(R_{i(j)})| \leq \sup_v \left| \frac{q(v | \mathbf{R})}{mI} - G(v) \right|$. \square

Lemma 3. *Under Assumption 1, we have as $I \rightarrow \infty$,*

$$\frac{1}{I} \sum_{i=1}^I \frac{q(R_{i1} | \mathbf{R})}{mI} \xrightarrow{a.s.} \mathbb{E}\{G(R_{i1})\}.$$

Proof. Note that

$$\frac{1}{I} \sum_{i=1}^I \frac{q(R_{i1} | \mathbf{R})}{mI} = \frac{1}{I} \sum_{i=1}^I G(R_{i1}) + \frac{1}{I} \sum_{i=1}^I \left\{ \frac{q(R_{i1} | \mathbf{R})}{mI} - G(R_{i1}) \right\}. \quad (2.10)$$

Since $G(R_{i1}), i = 1, 2, \dots$ are bounded and iid, by the law of large numbers, the first term in the RHS of (2.10) converges to $\mathbb{E}\{G(R_{i1})\}$ a.s.. By Lemma 1 and Lemma 2, the second term in the RHS of (2.10) converges to 0 a.s.. So the desired result

follows. □

Lemma 4. *Under Assumption 1, we have as $I \rightarrow \infty$,*

$$\frac{1}{I} \sum_{i=1}^I \frac{\bar{\mu}_i}{mI} \xrightarrow{a.s.} \mathbb{E} \left\{ \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \right\}.$$

Proof. Note that

$$\begin{aligned} & \left| \frac{\bar{\mu}_i}{mI} - \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \right| \\ &= \left| \max_{b \in [m-1]} \frac{\sum_{j=1}^b \frac{q(R_{i(j)}|\mathbf{R})}{mI} + \Gamma \sum_{j=b+1}^m \frac{q(R_{i(j)}|\mathbf{R})}{mI}}{b + \Gamma(m-b)} \right. \\ & \quad \left. - \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \right| \\ &\leq \max_{b \in [m-1]} \left| \frac{\sum_{j=1}^b \frac{q(R_{i(j)}|\mathbf{R})}{mI} + \Gamma \sum_{j=b+1}^m \frac{q(R_{i(j)}|\mathbf{R})}{mI}}{b + \Gamma(m-b)} - \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \right| \\ &\leq \max_{b \in [m-1]} \frac{\sum_{j=1}^b \left| \frac{q(R_{i(j)}|\mathbf{R})}{mI} - G(R_{i(j)}) \right| + \Gamma \sum_{j=b+1}^m \left| \frac{q(R_{i(j)}|\mathbf{R})}{mI} - G(R_{i(j)}) \right|}{b + \Gamma(m-b)}, \end{aligned}$$

together with Lemma 2, we have as $I \rightarrow \infty$,

$$\frac{\bar{\mu}_i}{mI} \xrightarrow{a.s.} \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)}. \quad (2.11)$$

Note that

$$\begin{aligned} \frac{1}{I} \sum_{i=1}^I \frac{\bar{\mu}_i}{mI} &= \frac{1}{I} \sum_{i=1}^I \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \\ &+ \frac{1}{I} \sum_{i=1}^I \left\{ \frac{\bar{\mu}_i}{mI} - \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \right\}. \quad (2.12) \end{aligned}$$

For the first term in the RHS of (2.12), note that $\max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)}$, $i = 1, 2, \dots$ are bounded iid random variables, by the strong law of large numbers, we have as $I \rightarrow \infty$,

$$\begin{aligned} & \frac{1}{I} \sum_{i=1}^I \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \\ & \xrightarrow{a.s.} \mathbb{E} \left\{ \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \right\}. \end{aligned}$$

The second term in the RHS of (2.12) converges to zero almost surely by (2.11). So the desired conclusion follows. \square

For simplicity, from now on, let

$$\varphi(\Gamma) = \mathbb{E} \left\{ \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \right\}.$$

Lemma 5. $\varphi(\Gamma)$ is continuous on $[1, +\infty)$.

Proof. For any $\Gamma_1, \Gamma_2 \in [1, +\infty)$,

$$\begin{aligned} & |\varphi(\Gamma_1) - \varphi(\Gamma_2)| \\ & \leq \mathbb{E} \left| \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_1 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_1(m-b)} \right. \\ & \quad \left. - \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_2 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_2(m-b)} \right| \\ & \leq \mathbb{E} \left\{ \max_{b \in [m-1]} \left| \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_1 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_1(m-b)} \right. \right. \\ & \quad \left. \left. - \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_2 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_2(m-b)} \right| \right\} \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E} \left\{ \max_{b \in [m-1]} \left| \left[\frac{\sum_{j=b+1}^m G(R_{i(j)})}{m-b} \right. \right. \right. \\
&\quad \left. \left. + \left\{ \sum_{j=1}^b G(R_{i(j)}) - \frac{b \cdot \sum_{j=b+1}^m G(R_{i(j)})}{m-b} \right\} \cdot \frac{1}{b + \Gamma_1(m-b)} \right] \right. \\
&\quad \left. - \left[\frac{\sum_{j=b+1}^m G(R_{i(j)})}{m-b} \right. \right. \\
&\quad \left. \left. + \left\{ \sum_{j=1}^b G(R_{i(j)}) - \frac{b \cdot \sum_{j=b+1}^m G(R_{i(j)})}{m-b} \right\} \cdot \frac{1}{b + \Gamma_2(m-b)} \right] \right| \Big\} \\
&= \mathbb{E} \left\{ \max_{b \in [m-1]} \left| \left(\frac{b}{b + \Gamma_1(m-b)} - \frac{b}{b + \Gamma_2(m-b)} \right) \right. \right. \\
&\quad \left. \left. \times \left(\frac{\sum_{j=1}^b G(R_{i(j)})}{b} - \frac{\sum_{j=b+1}^m G(R_{i(j)})}{m-b} \right) \right| \right\} \\
&\leq 2 \max_{b \in [m-1]} \left| \frac{b}{b + \Gamma_1(m-b)} - \frac{b}{b + \Gamma_2(m-b)} \right|, \quad (\text{since } G(R_{i(j)}) \leq 1)
\end{aligned}$$

then continuity of $\varphi(\Gamma)$ follows from the fact that $g_b(\Gamma) = \frac{b}{b + \Gamma(m-b)}$ is continuous on $[1, +\infty)$ for each $b \in [m-1]$. \square

Lemma 6. *Let X and Y be two random variables. Set $s_X = \sup\{t : \mathbb{P}(X \geq t) > 0\}$, $s_Y = \sup\{t : \mathbb{P}(Y \geq t) > 0\}$. Suppose that function $h : \mathbb{R} \rightarrow \mathbb{R}$ is continuously differentiable on $(c, +\infty)$, $h(c) = 0$, $h'(t) > 0$ for $t \in (c, s_X \vee s_Y)$, and $\mathbb{E}|h(X)\mathbb{1}_{X \geq c}| < \infty$, $\mathbb{E}|h(Y)\mathbb{1}_{Y \geq c}| < \infty$. If for any $t \in (c, s_X \vee s_Y)$, $\mathbb{P}(X \geq t) \leq \mathbb{P}(Y \geq t)$, we have $\mathbb{E}\{h(X)\mathbb{1}_{X \geq c}\} \leq \mathbb{E}\{h(Y)\mathbb{1}_{Y \geq c}\}$, and if there exists an open interval $\mathcal{I} \subset (c, s_X \vee s_Y)$ such that $\mathbb{P}(X \geq t) < \mathbb{P}(Y \geq t)$ for any $t \in \mathcal{I}$, we have $\mathbb{E}\{h(X)\mathbb{1}_{X \geq c}\} < \mathbb{E}\{h(Y)\mathbb{1}_{Y \geq c}\}$.*

Proof. We have

$$\begin{aligned}
\mathbb{E}\{h(X)\mathbb{1}_{X \geq c}\} &= \int_{\Omega} \int_c^X h'(t) \mathbb{1}_{X \geq c} dt d\mathbb{P} \quad (\text{since } h(c) = 0) \\
&= \int_{\Omega} \int_c^{+\infty} h'(t) \mathbb{1}_{X \geq t} dt d\mathbb{P}
\end{aligned}$$

$$\begin{aligned}
&= \int_c^{+\infty} \int_{\Omega} h'(t) \mathbb{1}_{X \geq t} d\mathbb{P} dt \quad (\text{by Fubini's theorem}) \\
&= \int_c^{+\infty} h'(t) \mathbb{P}(X \geq t) dt \\
&= \int_c^{s_X} h'(t) \mathbb{P}(X \geq t) dt.
\end{aligned}$$

Similarly, we have

$$\mathbb{E}\{h(Y) \mathbb{1}_{Y \geq c}\} = \int_c^{s_Y} h'(t) \mathbb{P}(Y \geq t) dt.$$

Note that if for any $t \in (c, s_X \vee s_Y)$, $\mathbb{P}(X \geq t) \leq \mathbb{P}(Y \geq t)$, then $s_X \leq s_Y$. So the desired conclusion follows immediately from the above two equalities and the assumption that $h'(t) > 0$ for $t \in (c, s_X \vee s_Y)$. \square

Lemma 7. *Under Assumptions 1-3, we have*

$$\frac{1}{m} \sum_{j=1}^m \mathbb{E}\{G(R_{ij})\} < \mathbb{E}\{G(R_{i1})\} < \mathbb{E}\{G(R_{i(m)})\}.$$

Proof. The second inequality follows immediately from applying Lemma 6 with $h(t) = G(t) = \frac{1}{m} \sum_{j=1}^m \max\{F_j(t) - F_j(c), 0\}$, $X = R_{i1}$ and $Y = R_{i(m)}$. For the first inequality, let $s_j = \sup\{t : \mathbb{P}(R_{ij} \geq t) > 0\}$. Assumption 3 implies that $s_1 = s = \max_j s_j$. Follow a similar calculation as in Lemma 6, by Assumption 3 we have

$$\begin{aligned}
\frac{1}{m} \sum_{j=1}^m \mathbb{E}\{G(R_{ij})\} &= \frac{1}{m} \sum_{j=1}^m \int_c^{s_j} G'(t) \mathbb{P}(R_{ij} \geq t) dt \\
&\leq \frac{1}{m} \sum_{j=1}^m \int_c^s G'(t) \mathbb{P}(R_{ij} \geq t) dt
\end{aligned}$$

$$\begin{aligned}
&= \int_c^s G'(t) \cdot \frac{1}{m} \sum_{j=1}^m \mathbb{P}(R_{ij} \geq t) dt \\
&< \int_c^s G'(t) \mathbb{P}(R_{i1} \geq t) dt \\
&= \mathbb{E}\{G(R_{i1})\}.
\end{aligned}$$

□

Lemma 8. *Under Assumptions 1-3, we have*

$$\lim_{\Gamma \rightarrow 1^+} \varphi(\Gamma) < \mathbb{E}\{G(R_{i1})\}, \quad \lim_{\Gamma \rightarrow +\infty} \varphi(\Gamma) > \mathbb{E}\{G(R_{i1})\}.$$

Proof. For any $\Gamma \geq 1$, since $G(R_{i(j)}) \leq 1$, we have

$$0 \leq \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma(m-b)} \leq 1.$$

By continuity of φ (Lemma 5) and bounded convergence theorem,

$$\begin{aligned}
\lim_{\Gamma \rightarrow 1^+} \varphi(\Gamma) &= \lim_{n \rightarrow \infty} \varphi\left(\frac{n+1}{n}\right) \\
&= \mathbb{E}\left\{ \lim_{n \rightarrow \infty} \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \frac{n+1}{n} \sum_{j=b+1}^m G(R_{i(j)})}{b + \frac{n+1}{n}(m-b)} \right\} \\
&= \frac{\sum_{j=1}^m \mathbb{E}\{G(R_{i(j)})\}}{m} \\
&= \frac{\sum_{j=1}^m \mathbb{E}\{G(R_{ij})\}}{m} \\
&< \mathbb{E}\{G(R_{i1})\}, \quad (\text{by Lemma 7})
\end{aligned}$$

$$\lim_{\Gamma \rightarrow +\infty} \varphi(\Gamma) = \lim_{n \rightarrow \infty} \varphi(n)$$

$$\begin{aligned}
&= \mathbb{E} \left\{ \lim_{n \rightarrow \infty} \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + n \sum_{j=b+1}^m G(R_{i(j)})}{b + n(m-b)} \right\} \\
&= \mathbb{E} \left\{ \max_{b \in [m-1]} \frac{\sum_{j=b+1}^m G(R_{i(j)})}{m-b} \right\} \\
&\geq \mathbb{E} \{ G(R_{i(m)}) \} \\
&> \mathbb{E} \{ G(R_{i1}) \}. \quad (\text{by Lemma 7})
\end{aligned}$$

□

Lemma 9. *Under Assumptions 1 and 2, $\varphi(\Gamma)$ is a strictly monotonically increasing function of Γ on $[1, +\infty)$.*

Proof. For any $b \in [m-1]$, and for any $1 \leq \Gamma_1 < \Gamma_2 < +\infty$, since

$$\frac{\sum_{j=1}^b G(R_{i(j)})}{b} - \frac{\sum_{j=b+1}^m G(R_{i(j)})}{m-b} \leq 0,$$

we have

$$\begin{aligned}
&\frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_2 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_2(m-b)} - \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_1 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_1(m-b)} \\
&= \frac{\sum_{j=b+1}^m G(R_{i(j)})}{m-b} + \left\{ \sum_{j=1}^b G(R_{i(j)}) - \frac{b \cdot \sum_{j=b+1}^m G(R_{i(j)})}{m-b} \right\} \cdot \frac{1}{b + \Gamma_2(m-b)} \\
&\quad - \frac{\sum_{j=b+1}^m G(R_{i(j)})}{m-b} - \left\{ \sum_{j=1}^b G(R_{i(j)}) - \frac{b \cdot \sum_{j=b+1}^m G(R_{i(j)})}{m-b} \right\} \cdot \frac{1}{b + \Gamma_1(m-b)} \\
&= b \left\{ \frac{\sum_{j=1}^b G(R_{i(j)})}{b} - \frac{\sum_{j=b+1}^m G(R_{i(j)})}{m-b} \right\} \left\{ \frac{1}{b + \Gamma_2(m-b)} - \frac{1}{b + \Gamma_1(m-b)} \right\} \\
&\geq 0,
\end{aligned}$$

where equality holds if and only if

$$\frac{\sum_{j=1}^b G(R_{i(j)})}{b} - \frac{\sum_{j=b+1}^m G(R_{i(j)})}{m-b} = 0.$$

Thus, we have

$$\begin{aligned} & \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_2 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_2(m-b)} \\ & \geq \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_1 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_1(m-b)}. \end{aligned}$$

That is, to show that $\varphi(\Gamma_2) > \varphi(\Gamma_1)$, it suffices to show that

$$\begin{aligned} & \mathbb{P} \left\{ \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_2 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_2(m-b)} \right. \\ & \quad \left. > \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_1 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_1(m-b)} \right\} > 0. \end{aligned}$$

We have

$$\begin{aligned} & \mathbb{P} \left\{ \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_2 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_2(m-b)} \right. \\ & \quad \left. > \max_{b \in [m-1]} \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_1 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_1(m-b)} \right\} \\ & \geq \mathbb{P} \left[\bigcap_{b=1}^{m-1} \left\{ \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_2 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_2(m-b)} \right. \right. \\ & \quad \left. \left. > \frac{\sum_{j=1}^b G(R_{i(j)}) + \Gamma_1 \sum_{j=b+1}^m G(R_{i(j)})}{b + \Gamma_1(m-b)} \right\} \right] \\ & = \mathbb{P} \left[\bigcap_{b=1}^{m-1} \left\{ \frac{\sum_{j=1}^b G(R_{i(j)})}{b} - \frac{\sum_{j=b+1}^m G(R_{i(j)})}{m-b} < 0 \right\} \right] \end{aligned}$$

$$\begin{aligned}
&\geq \mathbb{P}(R_{i(m)} > R_{i(m-1)} > \cdots > R_{i(1)} \geq c) \\
&= \mathbb{P}(R_{i(1)} \geq c) \quad (\text{by Assumption 1}) \\
&> 0, \quad (\text{by Assumption 2})
\end{aligned}$$

so the conclusion follows. \square

Theorem 1. By Lemma 5, Lemma 8 and Lemma 9, it is clear that equation $\varphi(\Gamma) = \mathbb{E}\{G(R_{i1})\}$ has a unique solution $\tilde{\Gamma}$ on $[1, +\infty)$. Note that,

$$\begin{aligned}
\Psi_{I,\Gamma} &= \mathbb{P}(T_{\text{abe}} \geq \xi_\alpha \mid \mathcal{Z}) \\
&= \mathbb{P}\left\{ \frac{\sum_{i=1}^I q(R_{i1} \mid \mathbf{R}) - \sum_{i=1}^I \bar{\mu}_i}{\sqrt{\sum_{i=1}^I \bar{v}_i}} \geq \Phi^{-1}(1 - \alpha) \right\} \\
&= \mathbb{P}\left\{ \frac{\sqrt{I} \left(\frac{1}{I} \sum_{i=1}^I \frac{q(R_{i1} \mid \mathbf{R})}{mI} - \frac{1}{I} \sum_{i=1}^I \frac{\bar{\mu}_i}{mI} \right)}{\sqrt{\frac{1}{I} \sum_{i=1}^I \frac{\bar{v}_i}{(mI)^2}}} \geq \Phi^{-1}(1 - \alpha) \right\}. \tag{2.13}
\end{aligned}$$

Since $q(R_{i(j)} \mid \mathbf{R}) \leq mI$, we have

$$\begin{aligned}
\frac{1}{I} \sum_{i=1}^I \frac{\bar{v}_i}{(mI)^2} &= \frac{1}{I} \sum_{i=1}^I \frac{\max_{b \in B_i} \bar{v}_{ib}}{(mI)^2} \\
&\leq \frac{1}{I} \sum_{i=1}^I \max_{b \in B_i} \frac{\sum_{j=1}^b \frac{q^2(R_{i(j)} \mid \mathbf{R})}{(mI)^2} + \Gamma \sum_{j=b+1}^m \frac{q^2(R_{i(j)} \mid \mathbf{R})}{(mI)^2}}{b + \Gamma(m - b)} \leq 1.
\end{aligned}$$

For $\Gamma < \tilde{\Gamma}$, by Lemma 9 we have $\varphi(\tilde{\Gamma}) > \varphi(\Gamma)$. Thus, by Lemma 3 and Lemma 4, as $I \rightarrow \infty$,

$$\frac{\sqrt{I} \left\{ \frac{1}{I} \sum_{i=1}^I \frac{q(R_{i1} \mid \mathbf{R})}{mI} - \frac{1}{I} \sum_{i=1}^I \frac{\bar{\mu}_i}{mI} \right\}}{\sqrt{\frac{1}{I} \sum_{i=1}^I \frac{\bar{v}_i}{(mI)^2}}} \simeq \frac{\sqrt{I} \{ \mathbb{E}(G(R_{i1})) - \varphi(\Gamma) \}}{\sqrt{\frac{1}{I} \sum_{i=1}^I \frac{\bar{v}_i}{(mI)^2}}}$$

$$\begin{aligned}
&= \frac{\sqrt{I}\{\varphi(\tilde{\Gamma}) - \varphi(\Gamma)\}}{\sqrt{\frac{1}{I}\sum_{i=1}^I \frac{\bar{v}_i}{(mI)^2}}} \\
&\geq \sqrt{I}\{\varphi(\tilde{\Gamma}) - \varphi(\Gamma)\} \\
&\rightarrow +\infty.
\end{aligned}$$

Similarly, we have for $\Gamma > \tilde{\Gamma}$,

$$\frac{\sqrt{I}\{\frac{1}{I}\sum_{i=1}^I \frac{q(R_{i1}|\mathbf{R})}{mI} - \frac{1}{I}\sum_{i=1}^I \frac{\bar{\mu}_i}{mI}\}}{\sqrt{\frac{1}{I}\sum_{i=1}^I \frac{\bar{v}_i}{(mI)^2}}} \rightarrow -\infty,$$

so the conclusion follows from (2.13). \square

Proof of Proposition 2

Proof. Suppose that $\mathbf{u}_0 \in \mathcal{U}$ is the unknown true vector of unmeasured confounders and $\Gamma_0 \leq \Gamma$ is the unknown true magnitude of hidden bias. Recall that ρ_Γ^* and y_Γ^* are the optimal values of (*) and (**) with sensitivity parameter Γ respectively. Since the constraint regions of (*) and (**) enlarge as Γ increases, we have $y_{\Gamma_0}^* \geq y_\Gamma^*$ and $\rho_{\Gamma_0}^* \geq \rho_\Gamma^*$. By Slepian's lemma, $\rho_{\Gamma_0}^* \geq \rho_\Gamma^*$ implies $Q_{\rho_{\Gamma_0}^*, \alpha} \leq Q_{\rho_\Gamma^*, \alpha}$. Thus we have

$$\begin{aligned}
&\mathbb{P}_{\Gamma_0, \mathbf{u}_0} \left(\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1,2\}} \frac{t_k - \mu_{k,\mathbf{u}}}{\sigma_{k,\mathbf{u}}} \geq Q_{\rho_\Gamma^*, \alpha} \mid \mathcal{F}, \mathcal{Z} \right) \\
&= \mathbb{P}_{\Gamma_0, \mathbf{u}_0} (y_\Gamma^* \geq Q_{\rho_\Gamma^*, \alpha} \mid \mathcal{F}, \mathcal{Z}) \\
&\leq \mathbb{P}_{\Gamma_0, \mathbf{u}_0} (y_{\Gamma_0}^* \geq Q_{\rho_{\Gamma_0}^*, \alpha} \mid \mathcal{F}, \mathcal{Z}) \\
&= \mathbb{P}_{\Gamma_0, \mathbf{u}_0} \left(\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1,2\}} \frac{t_k - \mu_{k,\mathbf{u}}}{\sigma_{k,\mathbf{u}}} \geq \max_{\mathbf{u} \in \mathcal{U}} Q_{\rho_{\mathbf{u}}, \alpha} \mid \mathcal{F}, \mathcal{Z}, \Gamma = \Gamma_0 \right)
\end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{P}_{\Gamma_0, \mathbf{u}_0} \left(\min_{\mathbf{u} \in \mathcal{U}} \left\{ \max_{k \in \{1,2\}} \frac{t_k - \mu_{k,\mathbf{u}}}{\sigma_{k,\mathbf{u}}} - Q_{\rho_{\mathbf{u}}, \alpha} \right\} \geq 0 \mid \mathcal{F}, \mathcal{Z}, \Gamma = \Gamma_0 \right) \\
&\leq \mathbb{P}_{\Gamma_0, \mathbf{u}_0} \left(\max_{k \in \{1,2\}} \frac{t_k - \mu_{k,\mathbf{u}_0}}{\sigma_{k,\mathbf{u}_0}} \geq Q_{\rho_{\mathbf{u}_0}, \alpha} \mid \mathcal{F}, \mathcal{Z}, \Gamma = \Gamma_0 \right) \\
&\rightarrow \alpha, \quad \text{as } I \rightarrow \infty
\end{aligned}$$

so the desired conclusion follows. \square

Proof of Theorem 2

Proof. Recall that using the Bonferroni adjustment to combine T_1 and T_2 takes the design sensitivity $\max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$; see [Rosenbaum \(2012\)](#). Since inequality (2.8) always holds, the test statistic used in testing procedure (2.9) implemented by Algorithm 1 uniformly dominates the one used by the Bonferroni adjustment, which implies $\tilde{\Gamma}_{1:2} \geq \max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$.

We then construct an example to show that $\tilde{\Gamma}_{1:2} > \max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$ is possible. That is, we show that the minimax procedure (developed in [Fogarty and Small \(2016\)](#)) implemented in the Step 2 of Algorithm 1 can result in improved design sensitivity by enforcing that unmeasured confounder must have the same impact on the probabilities of assignment to treatment for all scores in each component test on the same outcome variable.

Suppose we have I matched pairs and potential responses $r_{Tij}, r_{Cij} \in S = \{a + b\sqrt{2} : a \in \mathbb{Z}, b \in \mathbb{Z}\}$, $i = 1, \dots, I, j = 1, 2$. For $x \in S$, we define two functions $f_1(x) = a$ and $f_2(x) = b$ where $x = a + b\sqrt{2}$ for some $a, b \in \mathbb{Z}$. Note that both f_1 and f_2 are well-defined since for $a_1, b_1, a_2, b_2 \in \mathbb{Z}$, we have $a_1 + b_1\sqrt{2} = a_2 + b_2\sqrt{2}$ if and only if $a_1 = a_2$ and $b_1 = b_2$. Consider two test statistics $T_1 =$

$I^{-1} \sum_{i=1}^I D_i^{(1)}$ and $T_2 = I^{-1} \sum_{i=1}^I D_i^{(2)}$ where $D_i^{(1)} = (Z_{i1} - Z_{i2})(f_1(R_{i1}) - f_1(R_{i2}))$ and $D_i^{(2)} = (Z_{i1} - Z_{i2})(f_2(R_{i1}) - f_2(R_{i2}))$ are treated-minus-control paired difference of $f_1(R_{ij})$ and $f_2(R_{ij})$ respectively. Note that $r_{Tij} = f_1(r_{Tij}) + f_2(r_{Tij})\sqrt{2}$ and $r_{Cij} = f_1(r_{Cij}) + f_2(r_{Cij})\sqrt{2}$. Suppose that for all $i = 1, \dots, I, j = 1, 2$, the vector $(f_1(r_{Tij}), f_1(r_{Cij}), f_2(r_{Tij}), f_2(r_{Cij}))$ are i.i.d. realizations from the distribution:

$$(f_1(r_{Tij}), f_1(r_{Cij}), f_2(r_{Tij}), f_2(r_{Cij})) = \begin{cases} (3, 0, -1, 0) & \text{with probability } 1/2 \\ (-1, 0, 3, 0) & \text{with probability } 1/2. \end{cases}$$

Therefore, the vector of treated-minus-control paired differences $\mathbf{D}_i = (D_i^{(1)}, D_i^{(2)})$ are identically distributed as:

$$(D_i^{(1)}, D_i^{(2)}) = \begin{cases} (3, -1) & \text{with probability } 1/2 \\ (-1, 3) & \text{with probability } 1/2. \end{cases} \quad (2.14)$$

For $k = 1, 2$, let $\tilde{\Gamma}_k$ denote the design sensitivity for T_k under Fisher's sharp null of no treatment effect $H_0 : r_{Tij} = r_{Cij}, i = 1, \dots, I, j = 1, 2$, and let $\tilde{\Gamma}_{1,2}$ denote the design sensitivity for testing H_0 with T_1 and T_2 as the two component test statistics through the minimax procedure.

We at first show that $\tilde{\Gamma}_1 = \tilde{\Gamma}_2 = 3$. The design sensitivity is the value of $\tilde{\Gamma}$ such that the worst-case expectation $\bar{\mu}_{\Gamma,k}$ of T_k (i.e., the expectation of the limiting bounding distribution of T_k) with the magnitude of hidden bias $\Gamma = \tilde{\Gamma}$ equals the true expectation μ_k of T_k (i.e., the actual expectation of the limiting distribution of T_k) based on how the paired differences $D_i^{(k)}$ are generated, $k \in \{1, 2\}$; see [Rosenbaum \(2004\)](#). In this case, according to (2.14), it is clear that $\mu_1 = \mu_2 = 1$. To find the worst-case expectation of T_k at a given Γ , it is clear that under Fisher's sharp

null H_0 , any paired difference $D_i^{(k)}$ where a 3 was observed should be assigned probability $\Gamma/(1 + \Gamma)$ for the treated unit in the pair having the higher value of f_k . Similarly, when a -1 is observed the probability that the treated unit had the lower value of f_k should be set to $1/(1 + \Gamma)$. For any Γ , the worst-case expectation $\bar{\mu}_{\Gamma,k}$ of T_k is then:

$$\bar{\mu}_{\Gamma,k} = \frac{1}{2} \cdot 3 \cdot \left(\frac{\Gamma}{1+\Gamma} - \frac{1}{1+\Gamma} \right) + \frac{1}{2} \cdot (-1) \cdot \left(\frac{1}{1+\Gamma} - \frac{\Gamma}{1+\Gamma} \right) = \frac{2\Gamma - 2}{1 + \Gamma}.$$

To obtain $\tilde{\Gamma}_k$, we just need to solve the equation $\bar{\mu}_{\tilde{\Gamma}_k,1} = \mu_k$ with $\tilde{\Gamma}_k$, which, from the above arguments, can be written as: $(2\tilde{\Gamma}_k - 2)/(1 + \tilde{\Gamma}_k) = 1$. Thus, $\tilde{\Gamma}_k = 3$ for $k = 1, 2$.

We then show that $\tilde{\Gamma}_{1:2} = +\infty$. Note that asymptotically the minimax procedure fails to reject H_0 if the maximum over the unmeasured confounders of the minimum over the outcomes of the expectations of T_1 and T_2 under Fisher's sharp null H_0 exceeds the true expectation of the test statistic (in our example, 1 for both T_1 and T_2). Hence, the design sensitivity is the value of Γ such that the worst-case expectations under Fisher's sharp null for both test statistics exceed their true expectations. Asymptotically, of the I matched pairs $I/2$ will have the observed paired difference of $(3, -1)$ for their two score functions (f_1, f_2) and $I/2$ will have an observed paired difference of $(-1, 3)$. Separating the observed pairs into two sets according to their paired difference, let p_i be the probability that the treated individual receives the treatment in pair i in the $(3, -1)$ group, and let q_i be the probability that the treated individual receives the treatment in pair i of the $(-1, 3)$

group. For any given Γ , consider the following optimization problem:

$$\begin{aligned}
& \underset{y, p_i, q_i}{\text{maximize}} && y && (***) \\
& \text{subject to} && y \leq \frac{1}{I} \sum_{i=1}^{I/2} \{3 \cdot (2p_i - 1) + (-1) \cdot (2q_i - 1)\} \\
& && y \leq \frac{1}{I} \sum_{i=1}^{I/2} \{(-1) \cdot (2p_i - 1) + 3 \cdot (2q_i - 1)\} \\
& && \frac{1}{1+\Gamma} \leq p_i \leq \frac{\Gamma}{1+\Gamma} \quad i = 1, \dots, I/2 \\
& && \frac{1}{1+\Gamma} \leq q_i \leq \frac{\Gamma}{1+\Gamma} \quad i = 1, \dots, I/2
\end{aligned}$$

Let $\mathbf{x} = (y, p_1, q_1, \dots, p_{I/2}, q_{I/2})$, the above problem can be rewritten in canonical form:

$$\begin{aligned}
& \underset{y, p_i, q_i}{\text{maximize}} && f(\mathbf{x}) = y && (***) \\
& \text{subject to} && g_1(\mathbf{x}) = y - \frac{1}{I} \sum_{i=1}^{I/2} (6p_i - 2q_i) + 1 \leq 0 \\
& && g_2(\mathbf{x}) = y - \frac{1}{I} \sum_{i=1}^{I/2} (-2p_i + 6q_i) + 1 \leq 0 \\
& && s_{p_i}(\mathbf{x}) = p_i - \frac{\Gamma}{1+\Gamma} \leq 0 \quad i = 1, \dots, I/2 \\
& && s_{q_i}(\mathbf{x}) = q_i - \frac{\Gamma}{1+\Gamma} \leq 0 \quad i = 1, \dots, I/2 \\
& && t_{p_i}(\mathbf{x}) = -p_i + \frac{1}{1+\Gamma} \leq 0 \quad i = 1, \dots, I/2 \\
& && t_{q_i}(\mathbf{x}) = -q_i + \frac{1}{1+\Gamma} \leq 0. \quad i = 1, \dots, I/2
\end{aligned}$$

The above problem along with its canonical form considers the maximum over the unmeasured confounders of the minimum over the outcomes of the expectations of T_1 and T_2 under Fisher's sharp null H_0 . The design sensitivity would be the

value $\Gamma = \tilde{\Gamma}$ such that the optimal value y^* exceeds the true expectation 1. We claim that the optimal solution is $p_i^* = q_i^* = \Gamma/(1 + \Gamma)$ for each $i = 1, \dots, I/2$, yielding $y^* = (\Gamma - 1)/(1 + \Gamma)$. To show this, we proceed by showing that this solution satisfies the Karush–Kuhn–Tucker (KKT) conditions. Since both the objective functions and the constraints are affine, the KKT conditions are sufficient for proving optimality of a solution.

Associate KKT multipliers $\lambda_1, \lambda_2, \alpha_{pi}, \alpha_{qi}, \beta_{pi}, \beta_{qi}$ with the above constraints. Let $\lambda_1 = \lambda_2 = 1/2, \alpha_{pi} = \alpha_{qi} = 2/I, \beta_{pi} = \beta_{qi} = 0$ for all i . We just need to check the following four parts of KKT conditions hold:

(1) Stationarity: partial of the objective function equals sum of partials of constraints times their KKT multipliers for each variable.

$$y \quad 1 = \lambda_1 + \lambda_2 = 1/2 + 1/2$$

$$p_i \quad 0 = -\lambda_1 \cdot \frac{1}{I} \cdot 6 - \lambda_2 \cdot \frac{1}{I} \cdot (-2) + \alpha_{pi} - \beta_{pi} = -\frac{1}{2} \cdot \frac{1}{I} \cdot 6 - \frac{1}{2} \cdot \frac{1}{I} \cdot (-2) + \frac{2}{I} - 0$$

$$q_i \quad 0 = -\lambda_1 \cdot \frac{1}{I} \cdot (-2) - \lambda_2 \cdot \frac{1}{I} \cdot 6 + \alpha_{qi} - \beta_{qi} = -\frac{1}{2} \cdot \frac{1}{I} \cdot (-2) - \frac{1}{2} \cdot \frac{1}{I} \cdot 6 + \frac{2}{I} - 0.$$

(2) Primal feasibility: constraints must be satisfied. Let $\mathbf{x}^* =$

$$(y^*, p_1^*, q_1^*, \dots, p_{I/2}^*, q_{I/2}^*) = \left(\frac{\Gamma-1}{1+\Gamma}, \frac{\Gamma}{1+\Gamma}, \frac{\Gamma}{1+\Gamma}, \dots, \frac{\Gamma}{1+\Gamma}, \frac{\Gamma}{1+\Gamma} \right),$$

$$g_1(\mathbf{x}^*) = g_2(\mathbf{x}^*) = \frac{\Gamma-1}{1+\Gamma} - \frac{1}{I} \cdot \frac{I}{2} \cdot \frac{4\Gamma}{1+\Gamma} + 1 = 0,$$

and it is clear that $s_{pi}(\mathbf{x}^*), s_{qi}(\mathbf{x}^*), t_{pi}(\mathbf{x}^*), t_{qi}(\mathbf{x}^*) \leq 0$ are satisfied.

(3) Dual feasibility: KKT multipliers must be non-negative. This clearly holds based on our choices: $\lambda_1 = \lambda_2 = 1/2, \alpha_{pi} = \alpha_{qi} = 2/I, \beta_{pi} = \beta_{qi} = 0$ for each i .

(4) Complementary slackness: for each constraint, either the constraint is binding or the KKT multiplier is zero. This clearly holds since the only constraints which are not binding at the solution are $t_{pi}(\mathbf{x}^*)$ and $t_{qi}(\mathbf{x}^*)$. For these, $\beta_{pi} = \beta_{qi} = 0$.

Hence, the KKT conditions (1) – (4) are satisfied, which by KKT sufficiency implies optimality of the proposed solution $y^* = (\Gamma - 1)/(1 + \Gamma) < 1$ for any $1 \leq \Gamma < +\infty$. Hence, $\tilde{\Gamma}_{1:2} = +\infty$ since for any finite Γ , the optimal value y^* cannot exceed 1. Thus, the proof is complete. \square

We study the simulated power to illustrate the example with $\tilde{\Gamma}_{1:2} > \max\{\tilde{\Gamma}_1, \tilde{\Gamma}_2\}$ constructed in the proof of Theorem 2. We consider the test statistics T_1 and T_2 defined in the proof of Theorem 2. In this example constructed in the proof, we have $\tilde{\Gamma}_1 = \tilde{\Gamma}_2 = 3$ and $\tilde{\Gamma}_{1:2} = +\infty$. In Table 2.4, we report the simulated power of (i) using T_1 to test Fisher’s sharp null of no treatment effect H_0 , (ii) using T_2 to test H_0 , and (iii) using the minimax procedure to combine T_1 and T_2 to test H_0 , with level $\alpha = 0.05$, $\Gamma = 2, 2.5, 2.9, 3.1, 4, 6$ and sample size $I = 50, 100, 300$. The (one-sided) minimax procedure uses the critical value $\Phi^{-1}(1 - \alpha/2)$ as in [Fogarty and Small \(2016\)](#) in simulations. All the numbers are based on 10,000 replications.

Table 2.4: Simulated power of T_1 , T_2 and the minimax procedure combining T_1 and T_2 .

| Γ | T_1 | | | T_2 | | | Minimax | | |
|----------|----------|-----------|-----------|----------|-----------|-----------|----------|-----------|-----------|
| | $I = 50$ | $I = 100$ | $I = 300$ | $I = 50$ | $I = 100$ | $I = 300$ | $I = 50$ | $I = 100$ | $I = 300$ |
| 2.0 | 0.24 | 0.47 | 0.92 | 0.25 | 0.45 | 0.93 | 1.00 | 1.00 | 1.00 |
| 2.5 | 0.06 | 0.09 | 0.29 | 0.06 | 0.09 | 0.30 | 1.00 | 1.00 | 1.00 |
| 2.9 | 0.02 | 0.02 | 0.03 | 0.01 | 0.02 | 0.03 | 0.48 | 1.00 | 1.00 |
| 3.1 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.33 | 1.00 | 1.00 |
| 4.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.01 | 1.00 | 1.00 |
| 6.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 1.00 |

From Table 2.4, we can see that as sample size $I \rightarrow \infty$, the simulated power of all the three tests increases to 1 for $\Gamma < 3$. For $\Gamma > 3$, the simulated power of T_1 and T_2 is nearly zero, while the simulated power of using the minimax procedure to combine T_1 and T_2 in a sensitivity analysis still increases to 1 as I increases. This holds true even for Γ much larger than 3, which confirms that $\tilde{\Gamma}_1 = \tilde{\Gamma}_2 = 3$ and $\tilde{\Gamma}_{1:2} = +\infty$.

In the proof of Theorem 2, to find a case in which applying the minimax procedure results in the substantial gains in design sensitivity, we constructed an example with a perfect negative correlation between the two treated-minus-control paired differences of the two score functions f_1 and f_2 of the two tests T_1 and T_2 . In our example, the worst-case unmeasured confounder vector \mathbf{u} for one test is actually the best-case \mathbf{u} for the other test, and it is this conflict that yields the infinite design sensitivity. It is not necessary that the two test statistics be perfectly negatively correlated to attain an improved design sensitivity, and our procedure applied to independently distributed test statistics would also yield a design sensitivity that

is larger than the max of the component design sensitivities. As the correlation gets closer to 1, in each pair the worst-case \mathbf{u} for one test statistic is close to or exactly the worst-case \mathbf{u} for the other test statistic with higher and higher probability, which means that the gap of the max-min inequality (2.8) gets smaller. Recall that the gains in design sensitivity are resulted from the gap of the max-min inequality (2.8), we therefore expect the magnitude of the gains in design sensitivity gets larger as the correlation between the two test statistics gets closer to -1 , and gets smaller as the correlation gets closer to 1. This also explains why the gains in design sensitivity from applying our new adaptive testing procedure to combine the aberrant rank test and the Mantel-Haenszel test are small and hard to observe since these two component tests are highly positively correlated.

We study the simulated power to illustrate how the power of using the minimax procedure to combine T_1 and T_2 varies with the correlation between the two test statistics. In particular, we consider the following three settings of the joint distribution of two paired treated-minus-control differences $(D_i^{(1)}, D_i^{(2)})$:

- Setting 1 (perfectly positively correlated): $\mathbb{P}(D_i^{(1)} = 3, D_i^{(2)} = 3) = \mathbb{P}(D_i^{(1)} = -1, D_i^{(2)} = -1) = 1/2$
- Setting 2 (independent): $\mathbb{P}(D_i^{(1)} = 3, D_i^{(2)} = 3) = \mathbb{P}(D_i^{(1)} = 3, D_i^{(2)} = -1) = \mathbb{P}(D_i^{(1)} = -1, D_i^{(2)} = 3) = \mathbb{P}(D_i^{(1)} = -1, D_i^{(2)} = -1) = 1/4$
- Setting 3 (perfectly negatively correlated): $\mathbb{P}(D_i^{(1)} = 3, D_i^{(2)} = -1) = \mathbb{P}(D_i^{(1)} = -1, D_i^{(2)} = 3) = 1/2$.

Settings 1-3 have the same marginal distribution $\mathbb{P}(D_i^{(1)} = 3) = \mathbb{P}(D_i^{(2)} = 3) = \mathbb{P}(D_i^{(1)} = -1) = \mathbb{P}(D_i^{(2)} = -1) = 1/2$. That is, the power of using T_1 and T_2 to test H_0 is the same in each setting and has been reported in Table 2.4 if we still

set $\alpha = 0.05$, $\Gamma = 2, 2.5, 2.9, 3.1, 4, 6$ and sample size $I = 50, 100, 300$. Note that Setting 3 is exactly the example constructed in the proof. We set the critical value $= \Phi^{-1}(1 - \alpha/2)$ in each setting. All the numbers are based on 10,000 replications.

Table 2.5: Simulated power of the minimax procedure with the various correlations of the two component test statistics.

| Γ | Setting 1 | | | Setting 2 | | | Setting 3 | | |
|----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| | $I = 50$ | $I = 100$ | $I = 300$ | $I = 50$ | $I = 100$ | $I = 300$ | $I = 50$ | $I = 100$ | $I = 300$ |
| 2.0 | 0.10 | 0.31 | 0.87 | 0.38 | 0.86 | 1.00 | 1.00 | 1.00 | 1.00 |
| 2.5 | 0.01 | 0.05 | 0.17 | 0.11 | 0.46 | 0.99 | 1.00 | 1.00 | 1.00 |
| 2.9 | 0.00 | 0.00 | 0.01 | 0.03 | 0.19 | 0.85 | 0.48 | 1.00 | 1.00 |
| 3.1 | 0.00 | 0.00 | 0.00 | 0.02 | 0.11 | 0.67 | 0.33 | 1.00 | 1.00 |
| 4.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.01 | 1.00 | 1.00 |
| 6.0 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.20 | 1.00 |

From Table 2.5, we can see that the power of using the minimax procedure to combine T_1 and T_2 increases as the correlation of the two test statistics decreases, which agrees with the theoretical insight that the magnitude of the gains in design sensitivity gets larger as the correlation between the two test statistics gets closer to -1 . Also, Table 2.5 suggests that substantial gains in the design sensitivities can be expected in both Setting 2 (independent case) and Setting 3 (negatively correlated case). Especially, for Setting 2 (independent case), the power still increases as the sample size increases when $\Gamma = 4$, suggesting that the design sensitivity of the adaptive test (resulted from the minimax procedure) should be at least greater than or equal to 4 and is therefore evidently greater than the individual design sensitivity ($=3$). Therefore, substantial gains in design sensitivity resulted from the adaptive test can be evident with two negatively correlated, independent or

weakly positively correlated component test statistics.

Appendix B: Asymptotic Comparison With Correlation Within Strata

In Section 2.4.3, to make the simulations easier and clearer, we assume that R_{i1}, \dots, R_{im} are independent of each other within each stratum i . In practice, matching may introduce correlation within strata, so here we examine whether the pattern of asymptotic comparisons in Section 2.4.3 via the design sensitivity still holds with correlated outcomes within strata or not.

As in Section 2.4.2 and Section 2.4.3, without loss of generality, in Appendix B we still assume that $j = 1$ receives treatment and others receive control for each i , and $(R_{i1}, \dots, R_{im})^T = (r_{Ti1}, r_{Ci2}, \dots, r_{Cim})^T$ is an i.i.d. realization from a multivariate continuous distribution. Following Section 2.4.3, we still assume that $r_{Tij} = g(r_{Cij})$ for some deterministic function g . Instead of assuming r_{Ci1}, \dots, r_{Cim} are independent of each other within each stratum i as in Section 2.4.3, here we allow r_{Ci1}, \dots, r_{Cim} to be correlated with each other. Parallel with Models 1-4 in Section 2.4.3, we consider the following four models with correlated outcomes within strata:

- Model 5 (additive treatment effects, multivariate normal distribution with correlation): $r_{Tij} = r_{Cij} + \beta$, and $(r_{Ci1}, \dots, r_{Cim})$ follows a multivariate normal distribution with the mean vector being $(0, \dots, 0)$ and the covariance matrix being $\Sigma_{m \times m}$ where $\Sigma_{ii} = 1$ for $i = 1, \dots, m$ and $\Sigma_{ij} = 0.5$ for any $i \neq j$.
- Model 6 (additive treatment effects, multivariate Laplace distribution with correlation): $r_{Tij} = r_{Cij} + \beta$, and $(r_{Ci1}, \dots, r_{Cim})$ follows a multivariate Laplace distribution with the mean vector being $(0, \dots, 0)$ and the covari-

ance matrix being $\Sigma_{m \times m}$ where $\Sigma_{ii} = 1$ for $i = 1, \dots, m$ and $\Sigma_{ij} = 0.5$ for any $i \neq j$.

- Model 7 (multiplicative treatment effects, multivariate normal distribution with correlation): $r_{Tij} = \delta \cdot r_{Cij}$, and $(r_{C11}, \dots, r_{Cim})$ follows a multivariate normal distribution with the mean vector being $(0, \dots, 0)$ and the covariance matrix being $\Sigma_{m \times m}$ where $\Sigma_{ii} = 1$ for $i = 1, \dots, m$ and $\Sigma_{ij} = 0.5$ for any $i \neq j$.
- Model 8 (multiplicative treatment effects, multivariate Laplace distribution with correlation): $r_{Tij} = \delta \cdot r_{Cij}$, and $(r_{C11}, \dots, r_{Cim})$ follows a multivariate Laplace distribution with the mean vector being $(0, \dots, 0)$ and the covariance matrix being $\Sigma_{m \times m}$ where $\Sigma_{ii} = 1$ for $i = 1, \dots, m$ and $\Sigma_{ij} = 0.5$ for any $i \neq j$.

For Models 5-8, we still set the aberrant response threshold to be $c = 1$. Table 2.6 reports the design sensitivities of the Mantel-Haenszel test and the aberrant rank test under Models 5-8 with $m = 4$ (i.e., matching with three controls) and various β and δ . As described in Section 2.4.3, calculation of the design sensitivity can be done via Monte-Carlo simulations.

Table 2.6: Design sensitivities of the Mantel-Haenszel test and the aberrant rank test under Models 5-8 and matching with three controls with various parameters. The larger of the two design sensitivities of the two tests is in bold in each case.

| Model 5: additive, multivariate normal | | | |
|---|-----------------|-----------------|-----------------|
| Test statistic | $\beta = 0.50$ | $\beta = 0.75$ | $\beta = 1.00$ |
| Mantel-Haenszel | 3.47 | 6.45 | 11.99 |
| Aberrant rank | 3.98 | 7.70 | 14.74 |
| Model 6: additive, multivariate Laplace | | | |
| Test statistic | $\beta = 0.50$ | $\beta = 0.75$ | $\beta = 1.00$ |
| Mantel-Haenszel | 3.83 | 8.12 | 17.51 |
| Aberrant rank | 3.69 | 7.31 | 14.47 |
| Model 7: multiplicative, multivariate normal | | | |
| Test statistic | $\delta = 1.50$ | $\delta = 1.75$ | $\delta = 2.00$ |
| Mantel-Haenszel | 2.30 | 2.91 | 3.46 |
| Aberrant rank | 3.62 | 5.38 | 7.26 |
| Model 8: multiplicative, multivariate Laplace | | | |
| Test statistic | $\delta = 1.50$ | $\delta = 1.75$ | $\delta = 2.00$ |
| Mantel-Haenszel | 2.39 | 3.13 | 3.82 |
| Aberrant rank | 3.35 | 4.94 | 6.66 |

The pattern of Table 2.6 agrees with that of Table 2.1. First, the choice of the test statistic still has a huge influence on the design sensitivities in the presence of correlation within strata; see Models 7 and 8 in Table 2.6. Second, with correlated outcomes within strata, whether or not the aberrant rank test outperforms the Mantel-Haenszel test still depends on the unknown data generating process. From Table 2.6, we can see that under Models 5, 7 and 8, the aberrant rank test out-

performs the Mantel-Haenszel test in terms of design sensitivities. While under Model 6, the design sensitivities of the Mantel Haenszel test are larger than those of the aberrant rank test. Note that Models 5-8 are parallel with Models 1-4: both the marginal distributions of r_{Tij} and r_{Cij} are correspondingly equal for Models 1-4 versus Models 5-8. Therefore, the insights in Section 2.4.3 on when the aberrant rank test should be preferred over the Mantel-Haenszel test and other times the Mantel-Haenszel test should be preferred can still shed light on the simulation results in Table 2.6.

Appendix C: More Details on the Regularity Assumptions

We give a sufficient condition under which (T_1, T_2) considered in Section 2.5.2 is asymptotically bivariate normal in the sense that in large sample size, the distribution function of $\left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}}\right)$ can be approximated by that of $\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{\mathbf{u}} \\ \rho_{\mathbf{u}} & 1 \end{pmatrix}\right)$, where $\rho_{\mathbf{u}} = \mathbb{E}\left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}} \cdot \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \mid \mathcal{F}, \mathcal{Z}\right)$.

Let $\tilde{T}_{k,i} = (\sum_{j=1}^{n_i} Z_{ij}q_{ijk} - \sum_{j=1}^{n_i} p_{ij}q_{ijk})/\sigma_{k,\mathbf{u}}$ for $i = 1, \dots, I$ and $k \in \{1, 2\}$. Therefore, we have $\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}} = \sum_{i=1}^I \tilde{T}_{1,i}$ and $\frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} = \sum_{i=1}^I \tilde{T}_{2,i}$. Let $T_{\text{joint},i} = (\tilde{T}_{1,i}, \tilde{T}_{2,i})^T$ for $i = 1, \dots, I$. Proposition 3 gives a sufficient condition under which the desired asymptotic bivariate normality holds.

Proposition 3. We let $\Sigma_{\mathbf{u}} = \begin{pmatrix} 1 & \rho_{\mathbf{u}} \\ \rho_{\mathbf{u}} & 1 \end{pmatrix}$. Suppose that the following three assumptions hold: (i) Treatment assignments are independent across matched strata; (ii) As $I \rightarrow \infty$, $\Sigma_{\mathbf{u}}$ has a positive definite limit $\tilde{\Sigma}$; (iii) For any fixed nonzero vector $\lambda = (\lambda_1, \lambda_2)^T$, there exists

some $\delta > 0$, such that Lyapunov's condition

$$\lim_{I \rightarrow \infty} \frac{1}{(\lambda^T \Sigma_{\mathbf{u}} \lambda)^{1+\delta/2}} \sum_{i=1}^I \mathbb{E} \{ |\lambda^T T_{\text{joint},i}|^{2+\delta} \} = 0$$

is satisfied. Then we have as $I \rightarrow \infty$, (T_1, T_2) is asymptotically bivariate normal in the sense that

$$\Sigma_{\mathbf{u}}^{-1/2} \left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I_{2 \times 2}).$$

Proof. We first show that $\tilde{\Sigma}^{-1/2} \left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I_{2 \times 2})$. Through an application of the Cramér-Wold device (Billingsley, 1995) (Theorem 29.4), to ensure that $\tilde{\Sigma}^{-1/2} \left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I_{2 \times 2})$, we just need to ensure that for any nonzero vector $\tilde{\lambda} = (\tilde{\lambda}_1, \tilde{\lambda}_2)^T$, the following standardized deviate is asymptotically normal:

$$\frac{\tilde{\lambda}^T \tilde{\Sigma}^{-1/2} \left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T}{\sqrt{\tilde{\lambda}^T \tilde{\Sigma}^{-1} \tilde{\lambda}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \quad (2.15)$$

When treatment assignments are independent across matched strata (condition (i)), for each $I = 1, 2, \dots$, the sequence

$$A_I = \{ \tilde{\lambda}^T \tilde{\Sigma}^{-1/2} T_{\text{joint},1}, \dots, \tilde{\lambda}^T \tilde{\Sigma}^{-1/2} T_{\text{joint},I} \}$$

is a sequence of independent random variables, and the collection $\{A_1, A_2, \dots\}$ is a triangular array of random variables. Set $\lambda^T = \tilde{\lambda}^T \tilde{\Sigma}^{-1/2}$ in condition (iii) and apply Lyapunov central limit theorem (Billingsley, 1995) (Theorem 27.3) to $\{A_1, A_2, \dots\}$, we have

$$\frac{\tilde{\lambda}^T \tilde{\Sigma}^{-1/2} \left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T}{\sqrt{\tilde{\lambda}^T \tilde{\Sigma}^{-1/2} \Sigma_{\mathbf{u}} \tilde{\Sigma}^{-1/2} \tilde{\lambda}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Then (2.15) follows immediately from

$$\frac{\tilde{\lambda}^T \tilde{\Sigma}^{-1/2} \left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T}{\sqrt{\tilde{\lambda}^T \tilde{\lambda}}} = \frac{\tilde{\lambda}^T \tilde{\Sigma}^{-1/2} \left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T \sqrt{\tilde{\lambda}^T \tilde{\Sigma}^{-1/2} \Sigma_{\mathbf{u}} \tilde{\Sigma}^{-1/2} \tilde{\lambda}}}{\sqrt{\tilde{\lambda}^T \tilde{\Sigma}^{-1/2} \Sigma_{\mathbf{u}} \tilde{\Sigma}^{-1/2} \tilde{\lambda}}} \frac{\sqrt{\tilde{\lambda}^T \tilde{\Sigma}^{-1/2} \Sigma_{\mathbf{u}} \tilde{\Sigma}^{-1/2} \tilde{\lambda}}}{\sqrt{\tilde{\lambda}^T \tilde{\lambda}}}$$

$$\xrightarrow{\mathcal{L}} \mathcal{N}(0, 1) \text{ (by condition (ii) and Slutsky's theorem).}$$

So we have shown that $\tilde{\Sigma}^{-1/2} \left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I_{2 \times 2})$. Then note that

$$\begin{aligned} & \Sigma_{\mathbf{u}}^{-1/2} \left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T \\ &= \Sigma_{\mathbf{u}}^{-1/2} \tilde{\Sigma}^{1/2} \tilde{\Sigma}^{-1/2} \left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T \\ &\xrightarrow{\mathcal{L}} I_{2 \times 2} \cdot \mathcal{N}(\mathbf{0}, I_{2 \times 2}) \\ &\sim \mathcal{N}(\mathbf{0}, I_{2 \times 2}). \end{aligned}$$

That is, the distribution function of $\left(\frac{T_1 - \mu_{1,\mathbf{u}}}{\sigma_{1,\mathbf{u}}}, \frac{T_2 - \mu_{2,\mathbf{u}}}{\sigma_{2,\mathbf{u}}} \right)^T$ can be approximated by that of $\mathcal{N}(\mathbf{0}, \Sigma_{\mathbf{u}})$. So we are done. □

Appendix D: More Details on the Two-Stage Programming Method

In this section, we give a detailed derivation of the second stage of the two-stage programming method. Suppose that the prespecified level $\alpha < 1/2$. Once getting the rejection threshold $Q_{\rho_{\Gamma}^*, \alpha}$ through solving (*), we apply the method developed in [Fogarty and Small \(2016\)](#) to formulate the problem of checking if the inequality

$$\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1, 2\}} \frac{t_k - \mu_{k,\mathbf{u}}}{\sigma_{k,\mathbf{u}}} \geq Q_{\rho_{\Gamma}^*, \alpha}, \quad (Q_{\rho_{\Gamma}^*, \alpha} > 0 \text{ when } \alpha < 1/2) \quad (2.16)$$

would hold at a given Γ into checking if the optimal value of (**) is greater than or equal to zero or not.

For each fixed Γ , to check if (2.16) holds or not, we just need to check if

$$\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1,2\}} \left(t_k - \mu_{k,\mathbf{u}} - Q_{\rho_{\Gamma}^*, \alpha} \sigma_{k,\mathbf{u}} \right) \geq 0,$$

which can be transformed into solving the following optimization problem and checking if its optimal value ≥ 0 or not by introducing an auxiliary variable y :

$$\begin{aligned} & \underset{y, u_{ij}}{\text{minimize}} && y \\ & \text{subject to} && y \geq t_k - \mu_{k,\mathbf{u}} - Q_{\rho_{\Gamma}^*, \alpha} \sigma_{k,\mathbf{u}} \quad \forall k \in \{0, 1\} \\ & && 0 \leq u_{ij} \leq 1. \quad \forall i, j \end{aligned}$$

Note that the above constraints force y to be larger than $t_k - \mu_{k,\mathbf{u}} - Q_{\rho_{\Gamma}^*, \alpha} \sigma_{k,\mathbf{u}}$ for both $k = 1$ and $k = 2$, and drive us to search for the feasible value of $\mathbf{u} = (u_{11}, \dots, u_{In_1}) \in \mathcal{U}$ that allows for y to be as small as possible. This is a routine way of solving minimax problems ([Charalambous and Conn, 1978](#)). Therefore, the above optimization problem indeed seeks to find $\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1,2\}} (t_k - \mu_{k,\mathbf{u}} - Q_{\rho_{\Gamma}^*, \alpha} \sigma_{k,\mathbf{u}})$.

Recall that $w_{ij} = \exp(\gamma u_{ij})$ and $p_{ij} = w_{ij} / \sum_{j'=1}^{n_i} w_{ij'}$. Then the above optimization problem can be written as

$$\begin{aligned} & \underset{y, w_{ij}}{\text{minimize}} && y \\ & \text{subject to} && y \geq t_k - \mu_{k,\mathbf{u}} - Q_{\rho_{\Gamma}^*, \alpha} \sigma_{k,\mathbf{u}} \quad \forall k \in \{0, 1\} \\ & && 1 \leq w_{ij} \leq \Gamma, \quad \forall i, j \end{aligned}$$

where $\mu_{k,\mathbf{u}} = \sum_{i=1}^I \sum_{j=1}^{n_i} p_{ij} q_{ijk} = \sum_{i=1}^I \sum_{j=1}^{n_i} \frac{w_{ij}}{\sum_{j'=1}^{n_i} w_{ij'}} \cdot q_{ijk}$ and

$$\begin{aligned} \sigma_{k,\mathbf{u}} &= \sqrt{\sum_{i=1}^I \sum_{j=1}^{n_i} p_{ij} q_{ijk}^2 - \left(\sum_{i=1}^I \left(\sum_{j=1}^{n_i} p_{ij} q_{ijk} \right) \right)^2} \\ &= \sqrt{\sum_{i=1}^I \sum_{j=1}^{n_i} \frac{w_{ij}}{\sum_{j'=1}^{n_i} w_{ij'}} \cdot q_{ijk}^2 - \left(\sum_{i=1}^I \left(\sum_{j=1}^{n_i} \frac{w_{ij}}{\sum_{j'=1}^{n_i} w_{ij'}} \cdot q_{ijk} \right) \right)^2}. \end{aligned}$$

There are two terms that make the constraints of the above optimization problem complicated: the square root term $\sigma_{k,\mathbf{u}}$ and the linear-fractional term $w_{ij} / \sum_{j'=1}^{n_i} w_{ij'}$.

We first consider how to get rid of the square root term $\sigma_{k,\mathbf{u}}$. Recall that we are just concerned about if the optimal value of the above optimization problem ≥ 0 or not. We introduce a prespecified sufficiently large constant 'M' and two axillary variables b_1 and b_2 , and then instead consider the following adjusted optimization problem:

$$\begin{aligned} &\underset{y, w_{ij}, b_k}{\text{minimize}} && y \\ &\text{subject to} && y \geq (t_k - \mu_{k,\mathbf{u}})^2 - Q_{\rho_\Gamma^*, \alpha}^2 \sigma_{k,\mathbf{u}}^2 - M b_k \quad \forall k \in \{0, 1\} \\ &&& 1 \leq w_{ij} \leq \Gamma \quad \forall i, j \\ &&& b_k \in \{0, 1\} \quad \forall k \in \{0, 1\} \\ &&& -M b_k \leq t_k - \mu_{k,\mathbf{u}} \leq M(1 - b_k). \quad \forall k \in \{0, 1\} \end{aligned}$$

Note that when M is sufficiently large, for all $k \in \{0, 1\}$ and $1 \leq w_{ij} \leq \Gamma$, we have $-M b_k \leq t_k - \mu_{k,\mathbf{u}} \leq M(1 - b_k)$ for either $b_k = 0$ or $b_k = 1$. When $b_k = 0$, we have the constraints $0 \leq t_k - \mu_{k,\mathbf{u}} \leq M$ and $y \geq (t_k - \mu_{k,\mathbf{u}})^2 - Q_{\rho_\Gamma^*, \alpha}^2 \sigma_{k,\mathbf{u}}^2$. When M is sufficiently large, for any $\mathbf{u} \in [0, 1]^N$ (or equivalently, for any $(w_{11}, \dots, w_{I n_I}) \in [1, \Gamma]^N$) such that $0 \leq t_k - \mu_{k,\mathbf{u}} \leq M$, we have $(t_k - \mu_{k,\mathbf{u}})^2 - Q_{\rho_\Gamma^*, \alpha}^2 \sigma_{k,\mathbf{u}}^2 \geq 0$ if and only if

$t_k - \mu_{k,\mathbf{u}} - Q_{\rho_{\Gamma}^*, \alpha} \sigma_{k,\mathbf{u}} \geq 0$. When $b_k = 1$, we have the constraints $-M \leq t_k - \mu_{k,\mathbf{u}} \leq 0$ and $y \geq (t_k - \mu_{k,\mathbf{u}})^2 - Q_{\rho_{\Gamma}^*, \alpha}^2 \sigma_{k,\mathbf{u}}^2 - M$. When M is sufficiently large, for any $\mathbf{u}_1, \mathbf{u}_2 \in [0, 1]^N$, we have $(t_k - \mu_{k,\mathbf{u}_2})^2 - Q_{\rho_{\Gamma}^*, \alpha}^2 \sigma_{k,\mathbf{u}_2}^2 > (t_k - \mu_{k,\mathbf{u}_1})^2 - Q_{\rho_{\Gamma}^*, \alpha}^2 \sigma_{k,\mathbf{u}_1}^2 - M$ and $(t_k - \mu_{k,\mathbf{u}_1})^2 - Q_{\rho_{\Gamma}^*, \alpha}^2 \sigma_{k,\mathbf{u}_1}^2 - M < 0$. Therefore, the above ‘ M ’ constraint imposes a directional error to ensure that we will not falsely reject the null if evidence pointed in the opposite direction of the alternative, i.e., the cases in which $\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1,2\}} (t_k - \mu_{k,\mathbf{u}})^2 / \sigma_{k,\mathbf{u}}^2 \geq Q_{\rho_{\Gamma}^*, \alpha}^2$ while $\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1,2\}} (t_k - \mu_{k,\mathbf{u}}) / \sigma_{k,\mathbf{u}} < Q_{\rho_{\Gamma}^*, \alpha}$.

We then consider how to get rid of the linear-fractional term $w_{ij} / \sum_{j'=1}^{n_i} w_{ij'}$. A routine way of transforming a linear-fractional term into linear terms is through applying the Charnes-Cooper transformation (Charnes and Cooper, 1962):

$$p_{ij} = \frac{w_{ij}}{\sum_{j'=1}^{n_i} w_{ij'}} = \frac{\exp(\gamma u_{ij})}{\sum_{j'=1}^{n_i} \exp(\gamma u_{ij'})}, \quad s_i = \frac{1}{\sum_{j'=1}^{n_i} w_{ij'}} = \frac{1}{\sum_{j'=1}^{n_i} \exp(\gamma u_{ij'})}.$$

Then the above optimization problem can be transformed into the following quadratically constrained linear program as stated in section 2.5.2:

$$\begin{aligned} & \text{minimize } y \quad (**) \\ & \quad \quad \quad y, p_{ij}, s_i, b_k \\ & \text{subject to } y \geq (t_k - \mu_{k,\mathbf{u}})^2 - Q_{\rho_{\Gamma}^*, \alpha}^2 \sigma_{k,\mathbf{u}}^2 - M b_k \quad \forall k \in \{0, 1\} \\ & \quad \quad \quad \sum_{j=1}^{n_i} p_{ij} = 1 \quad \forall i \\ & \quad \quad \quad s_i \leq p_{ij} \leq \Gamma s_i \quad \forall i, j \\ & \quad \quad \quad p_{ij} \geq 0 \quad \forall i, j \\ & \quad \quad \quad b_k \in \{0, 1\} \quad \forall k \in \{0, 1\} \\ & \quad \quad \quad -M b_k \leq t_k - \mu_{k,\mathbf{u}} \leq M(1 - b_k). \quad \forall k \in \{0, 1\} \end{aligned}$$

Therefore, to judge if (2.16) holds or not, we just need to check whether the optimal

value y_1^* of (**) satisfies $y_1^* \geq 0$. If it is, we reject the null; otherwise, we fail to reject.

Appendix E: Adaptive Approach in Full Matching Case

In this section, we discuss how the adaptive approach developed in the main text can be easily adjusted to allow for full matching case, i.e., the constraint that $\sum_{j=1}^{n_i} Z_{ij} \in \{1, n_i - 1\}$ for all $i = 1, \dots, I$ (i.e., either one treated individual and one or more controls, or one control and one or more treated individuals, within each stratum). Let $I = I_1 + I_2$. We rearrange the index i of each stratum such that $\sum_{j=1}^{n_i} Z_{ij} = 1$ for $i = 1, \dots, I_1$ and $\sum_{j=1}^{n_i} Z_{ij} = n_i - 1$ for $i = I_1 + 1, \dots, I_1 + I_2$. Let $\tilde{\mathcal{Z}}$ be the collection of the treatment assignment indicators $\mathbf{Z} = (Z_{11}, \dots, Z_{In_I})$ such that $\mathbf{Z} \in \tilde{\mathcal{Z}}$ if and only if $\sum_{j=1}^{n_i} Z_{ij} = 1$ for all $i = 1, \dots, I_1$ and $\sum_{j=1}^{n_i} Z_{ij} = n_i - 1$ for all $i = I_1 + 1, \dots, I_1 + I_2$. Let $p_{ij} = \mathbb{P}(Z_{ij} = 1 \mid \mathcal{F}, \tilde{\mathcal{Z}})$, $\tilde{p}_{ij} = \mathbb{P}(Z_{ij} = 0 \mid \mathcal{F}, \tilde{\mathcal{Z}}) = 1 - p_{ij}$ and $T_{k,i} = \sum_{j=1}^{n_i} Z_{ij} q_{ijk}$ for $i = 1, \dots, I, j = 1, \dots, n_i, k = 1, 2$. Therefore, we have $T_1 = \sum_{i=1}^I T_{1,i}$ and $T_2 = \sum_{i=1}^I T_{2,i}$. In this case, we have

$$\begin{cases} p_{ij} = \exp(\gamma u_{ij}) / \sum_{j'=1}^{n_i} \exp(\gamma u_{ij'}) & \text{for } i = 1, \dots, I_1, j = 1, \dots, n_i, \\ \tilde{p}_{ij} = \exp(-\gamma u_{ij}) / \sum_{j'=1}^{n_i} \exp(-\gamma u_{ij'}) & \text{for } i = I_1 + 1, \dots, I_1 + I_2, j = 1, \dots, n_i. \end{cases}$$

Then we have

$$\begin{aligned}
\tilde{\mu}_{k,\mathbf{u}} &= \mathbb{E}_{\Gamma,\mathbf{u}}(T_k \mid \mathcal{F}, \tilde{\mathcal{Z}}) = \sum_{i=1}^{I_1} \sum_{j=1}^{n_i} p_{ij} q_{ijk} + \sum_{i=I_1+1}^{I_1+I_2} \sum_{j=1}^{n_i} q_{ijk} - \sum_{i=I_1+1}^{I_1+I_2} \sum_{j=1}^{n_i} \tilde{p}_{ij} q_{ijk}, \\
\tilde{\sigma}_{k,\mathbf{u}}^2 &= \text{Var}_{\Gamma,\mathbf{u}}(T_k \mid \mathcal{F}, \tilde{\mathcal{Z}}) = \sum_{i=1}^{I_1} \sum_{j=1}^{n_i} p_{ij} q_{ijk}^2 - \sum_{i=1}^{I_1} \left(\sum_{j=1}^{n_i} p_{ij} q_{ijk} \right)^2 \\
&\quad + \sum_{i=I_1+1}^{I_1+I_2} \sum_{j=1}^{n_i} \tilde{p}_{ij} q_{ijk}^2 - \sum_{i=I_1+1}^{I_1+I_2} \left(\sum_{j=1}^{n_i} \tilde{p}_{ij} q_{ijk} \right)^2, \\
\text{Cov}_{\Gamma,\mathbf{u}}(T_1, T_2 \mid \mathcal{F}, \tilde{\mathcal{Z}}) &= \sum_{i=1}^{I_1} \sum_{j=1}^{n_i} p_{ij} q_{ij1} q_{ij2} - \sum_{i=1}^{I_1} \left(\sum_{j=1}^{n_i} p_{ij} q_{ij1} \right) \left(\sum_{j=1}^{n_i} p_{ij} q_{ij2} \right) \\
&\quad + \sum_{i=I_1+1}^{I_1+I_2} \sum_{j=1}^{n_i} \tilde{p}_{ij} q_{ij1} q_{ij2} - \sum_{i=I_1+1}^{I_1+I_2} \left(\sum_{j=1}^{n_i} \tilde{p}_{ij} q_{ij1} \right) \left(\sum_{j=1}^{n_i} \tilde{p}_{ij} q_{ij2} \right), \\
\tilde{\rho}_{\mathbf{u}} &= \mathbb{E} \left(\frac{T_1 - \tilde{\mu}_{1,\mathbf{u}}}{\tilde{\sigma}_{1,\mathbf{u}}} \cdot \frac{T_2 - \tilde{\mu}_{2,\mathbf{u}}}{\tilde{\sigma}_{2,\mathbf{u}}} \mid \mathcal{F}, \tilde{\mathcal{Z}} \right) = \frac{\text{Cov}_{\Gamma,\mathbf{u}}(T_1, T_2 \mid \mathcal{F}, \tilde{\mathcal{Z}})}{\tilde{\sigma}_{1,\mathbf{u}} \tilde{\sigma}_{2,\mathbf{u}}}. \tag{2.17}
\end{aligned}$$

Similar to the main text, we let $w_{ij} = \exp(\gamma u_{ij})$. We further let $\tilde{w}_{ij} = \exp(-\gamma u_{ij}) = w_{ij}^{-1}$. In the first stage, to find the worst-case correlation $\min_{\mathbf{u} \in \mathcal{U}} \tilde{\rho}_{\mathbf{u}}$, we just need to solve the following optimization problem which just slightly adjusts the box constraints of the optimization problem (*):

$$\begin{aligned}
&\underset{w_{ij}, \tilde{w}_{ij}}{\text{minimize}} \quad \tilde{\rho}_{\mathbf{u}} \quad (\diamond) \\
&\text{subject to} \quad 1 \leq w_{ij} \leq \Gamma \quad i = 1, \dots, I_1, j = 1, \dots, n_i \\
&\quad \Gamma^{-1} \leq \tilde{w}_{ij} \leq 1, \quad i = I_1 + 1, \dots, I_1 + I_2, j = 1, \dots, n_i
\end{aligned}$$

where $\rho_{\mathbf{u}}$ is as defined in (2.17) with

$$\begin{cases} p_{ij} = w_{ij} / \sum_{j'=1}^{n_i} w_{ij'} & \text{for } i = 1, \dots, I_1, j = 1, \dots, n_i, \\ \tilde{p}_{ij} = \tilde{w}_{ij} / \sum_{j'=1}^{n_i} \tilde{w}_{ij'} & \text{for } i = I_1 + 1, \dots, I_1 + I_2, j = 1, \dots, n_i. \end{cases}$$

Similar to (*) in the main text, the optimization problem (\diamond) can be solved approximately in a reasonable amount of time by the L-BFGS-B algorithm (Byrd et al., 1995; Zhu et al., 1997). Denote the optimal value of (\diamond) with sensitivity parameter Γ as ρ_Γ^\diamond . Then the corresponding worst-case quantile $\max_{\mathbf{u} \in \mathcal{U}} Q_{\tilde{\rho}_{\mathbf{u}, \alpha}}$ equals $Q_{\rho_\Gamma^\diamond, \alpha}$ by Slepian's lemma.

As discussed in the main text, to determine if we should reject the null with level α and a given Γ in a sensitivity analysis, we then need to check if $\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1, 2\}} (t_k - \tilde{\mu}_{k, \mathbf{u}}) / \tilde{\sigma}_{k, \mathbf{u}} \geq Q_{\rho_\Gamma^\diamond, \alpha}$ at that given Γ . Adapting a similar argument to that in Appendix D, this procedure can be implemented through setting

$$\begin{cases} s_i = 1 / \sum_{j'=1}^{n_i} \exp(\gamma u_{ij'}) & \text{for } i = 1, \dots, I_1 \\ \tilde{s}_i = 1 / \sum_{j'=1}^{n_i} \exp(-\gamma u_{ij'}) & \text{for } i = I_1 + 1, \dots, I_1 + I_2. \end{cases}$$

and solving the following quadratically constrained linear program with M being

a sufficiently large constant:

$$\begin{aligned}
& \underset{y, p_{ij}, \tilde{p}_{ij}, s_i, \tilde{s}_i, b_k}{\text{minimize}} && y \quad (\diamond\diamond) \\
& \text{subject to} && y \geq (t_k - \tilde{\mu}_{k,\mathbf{u}})^2 - Q_{\rho_{\Gamma}^{\diamond}, \alpha}^2 \tilde{\sigma}_{k,\mathbf{u}}^2 - Mb_k \quad \forall k \in \{0, 1\} \\
& && \sum_{j=1}^{n_i} p_{ij} = 1 \quad \forall i = 1, \dots, I_1, \\
& && \sum_{j=1}^{n_i} \tilde{p}_{ij} = 1 \quad \forall i = I_1 + 1, \dots, I_1 + I_2, \\
& && s_i \leq p_{ij} \leq \Gamma s_i \quad \forall i = 1, \dots, I_1, j = 1, \dots, n_i, \\
& && \Gamma^{-1} \tilde{s}_i \leq \tilde{p}_{ij} \leq \tilde{s}_i \quad \forall i = I_1 + 1, \dots, I_1 + I_2, j = 1, \dots, n_i, \\
& && p_{ij} \geq 0 \quad \forall i = 1, \dots, I_1, j = 1, \dots, n_i, \\
& && \tilde{p}_{ij} \geq 0 \quad \forall i = I_1 + 1, \dots, I_1 + I_2, j = 1, \dots, n_i, \\
& && b_k \in \{0, 1\} \quad \forall k \in \{0, 1\} \\
& && -Mb_k \leq t_k - \tilde{\mu}_{k,\mathbf{u}} \leq M(1 - b_k), \quad \forall k \in \{0, 1\}
\end{aligned}$$

and checking whether the optimal value $y_{\Gamma}^{\diamond} \geq 0$.

Algorithm 2: Two-stage programming procedure in full matching case

Step 0: Re-order the matched strata such that with $I = I_1 + I_2$, we have

$$\sum_{j=1}^{n_i} Z_{ij} = 1 \text{ for } i = 1, \dots, I_1 \text{ and } \sum_{j=1}^{n_i} Z_{ij} = n_i - 1 \text{ for } i = I_1 + 1, \dots, I_1 + I_2;$$

Input: Sensitivity parameter Γ ; level α of the one-sided test; treatment

assignment indicator vector $\mathbf{Z} = (Z_{11}, \dots, Z_{I_1 I_1})^T$; the score vector

$\mathbf{q}_1 = (q_{111}, \dots, q_{I_1 I_1})^T$ associated with $T_1 = \sum_{i=1}^I \sum_{j=1}^{n_i} Z_{ij} q_{ij1}$; the score vector

$\mathbf{q}_2 = (q_{112}, \dots, q_{I_1 I_1})^T$ associated with $T_2 = \sum_{i=1}^I \sum_{j=1}^{n_i} Z_{ij} q_{ij2}$;

Step 1: Solve (\diamond) to get the worst-case correlation ρ_{Γ}^{\diamond} along with the

corresponding worst-case quantile $Q_{\rho_{\Gamma}^{\diamond}, \alpha}$;

Step 2: Solve $(\diamond\diamond)$ with $Q_{\rho_{\Gamma}^{\diamond}, \alpha}$ obtained from Step 1, and get the corresponding

optimal value y_{Γ}^{\diamond} ;

Output: If $y_{\Gamma}^{\diamond} \geq 0$, we reject the null; otherwise, we fail to reject.

Appendix F: Simulated Size of a Sensitivity Analysis

We study the simulated size of the Mantel-Haenszel test, the aberrant rank test and the adaptive test implementing Algorithm 1 with the above two tests as the component tests under the aberrant null for various Γ . Specifically, we set $\beta = 0$ in Models 1 and 2 or $\delta = 1$ in Models 3 and 4. We set $\alpha = 0.05$, $c = 1$ and $m = 4$ (matching with three controls), and as in Models 1-4, we consider two cases: either F is a standard normal distribution or a standard Laplace distribution. Both $I = 100$ and $I = 1000$ matched strata are considered. Each simulated size of the Mantel-Haenszel test and the aberrant rank test is based on 20,000 replications and each simulated size of the adaptive test is based on 2,000 replications. The simulation results are presented in Table 2.7.

Table 2.7: Simulated size of the Mantel-Haenszel test, the aberrant rank test and the adaptive test implementing Algorithm 1 with the above two tests as the component tests under the aberrant null. We set $\alpha = 0.05$, $c = 1$ and $m = 4$ (matching with three controls).

| Normal | $I = 100$ Matched Strata | | | $I = 1000$ Matched Strata | | |
|-----------------|--------------------------|----------|----------|---------------------------|----------|----------|
| | M-H | Aberrant | Adaptive | M-H | Aberrant | Adaptive |
| $\Gamma = 1.00$ | 0.051 | 0.054 | 0.045 | 0.051 | 0.053 | 0.042 |
| $\Gamma = 1.05$ | 0.037 | 0.040 | 0.044 | 0.018 | 0.021 | 0.020 |
| $\Gamma = 1.10$ | 0.027 | 0.031 | 0.035 | 0.005 | 0.007 | 0.003 |
| Laplace | $I = 100$ Matched Strata | | | $I = 1000$ Matched Strata | | |
| | M-H | Aberrant | Adaptive | M-H | Aberrant | Adaptive |
| $\Gamma = 1.00$ | 0.053 | 0.055 | 0.044 | 0.049 | 0.053 | 0.042 |
| $\Gamma = 1.05$ | 0.039 | 0.043 | 0.041 | 0.020 | 0.021 | 0.024 |
| $\Gamma = 1.10$ | 0.029 | 0.032 | 0.028 | 0.007 | 0.009 | 0.009 |

We here provide some insights into the simulation results presented in Table 2.7. Following the previous literature (e.g., [Imbens and Rosenbaum, 2005](#); [Heng et al., 2020](#)), the simulated size of each test in each scenario is calculated under the situation when, parallel with the favorable situation, there is no treatment effect and no hidden bias. When $\Gamma = 1$, we can see that all three tests can approximately preserve a 0.05 type I error rate control with realistic sample sizes. When $\Gamma > 1$, each simulated size of a sensitivity analysis with that prespecified sensitivity parameter Γ is less than 0.05 for all three tests, and decreases as the prespecified sensitivity parameter Γ increases. This pattern agrees with that of the power of a sensitivity analysis as shown in Table 2.2. This is because it is more and more improbable that a sensitivity analysis conducted at a larger and larger Γ will reject, either correctly

or falsely, the null hypothesis of no treatment effect if, in fact, the treatment assignment is random. This is a general pattern that is not only shared by the above three tests, but also most of the plausible non-parametric tests used in matched observational studies (e.g., [Heng et al., 2020](#)) (Section 3.5). Note that for all three tests, the simulated size drops substantially as Γ increases when the sample size is relatively large ($I = 1000$). This is also expected since when the sample size is large, the standard error of a test statistic should be relatively small compared to the magnitude of bias and in this case the size of a test is driven more by the bias. Another way to understand this pattern is from the design sensitivity. Recall that when there is an actual treatment effect, the chance of rejecting the null hypothesis of no treatment effect (i.e., the power of a test) in a sensitivity analysis conducted with $\Gamma > \tilde{\Gamma}$ goes to zero as the sample size I goes to infinity, where the design sensitivity $\tilde{\Gamma}$ approaches 1 as the actual treatment effect approaches zero. Therefore, if there is in fact no treatment effect, the chance of rejecting the null hypothesis (i.e., the size of a test) in a sensitivity analysis with any $\Gamma > 1$ goes to zero as the sample size increases.

Appendix G: More Details on Sections 2.1 and 2.7

In Section 2.1.1, we have mentioned that: “Numerous causal problems share a similar structure with that of the causal determinants of malnutrition, where we care about whether a certain treatment would change the pattern of some aberrant response (e.g., stunted growth) rather than the average treatment effect over the whole population.” We here provide two more examples of such type of causal problems.

- According to [WHO \(2008b\)](#), anemia in adult men can be defined as blood hemoglobin (Hb) concentrations < 130 g / l, and related studies typically fo-

cus on the prevalence and severity of anemia, instead of the change of average blood Hb concentrations among the whole population; see [Adamu et al. \(2017\)](#).

- A commonly used definition for low birth weight is infant's weight at birth being less than or equal to 2500 g; see [Kramer \(1987\)](#). Related studies are typically concerned about the low birth weight rate and the severity of low birth weight among the study population, instead of the change of average birth weight among that study population; see [Paneth \(1995\)](#) and [Schieve et al. \(2002\)](#).

In Section 2.1.1.1, we examine the causal problem of the potential effect of teenage pregnancy on stunting with children's level data from the Kenya 2003 Demographic and Health Surveys (DHS), which is available at Integrated Public Use Microdata Series (IPUMS). We here provide some motivating examples from the previous literature. According to [Darteh et al. \(2014\)](#), a causal effect of teenage pregnancy on stunting could arise "as a result of the fact that young mothers require adequate nutrition to fully grow into adults; thus, they struggle with their children over the little food the mother eats." [Van de Poel et al. \(2007\)](#) argued that "Children of younger mothers could be more prone to malnutrition because of physiological immaturity and social and psychological stress that come with child bearing at young age." We also summarize the total number of stunting cases among the treated individuals and controls in the matched data in Table 2.8.

Table 2.8: The total number of stunting cases among the treated individuals and controls in the matched data.

| | Stunted | Non-stunted | Total | Percentage |
|---------|---------|-------------|-------|------------|
| Treated | 55 | 95 | 150 | 36.7% |
| Control | 125 | 325 | 450 | 27.8% |
| Total | 180 | 420 | 600 | 30.0% |

In Section 2.7, we report the worst-case p-values of the Mantel-Haenszel test, the aberrant rank test and the adaptive test. The worst-case p-values of the Mantel-Haenszel test and the aberrant rank test are obtained from the results of the asymptotic approximation of the worst-case p-value in Section 2.3 and Section 2.4.1. Especially, for the aberrant rank test, we apply the asymptotic separability algorithm proposed in [Gastwirth et al. \(2000\)](#) to find an approximate worst-case p-value under some mild conditions on the response vector and the treatment assignment probabilities. See Proposition 1 in [Gastwirth et al. \(2000\)](#) for more details about the mild conditions under which the asymptotic separability algorithm is applicable. Note that the adaptive testing procedure (2.9) is a procedure that directly determines if we should reject the null hypothesis or not and does not directly involve the worst-case p-value in the traditional sense. Instead, the worst-case p-value reported by the adaptive test in Table 2.3 is the value of the prespecified α such that the adaptive testing procedure (2.9) barely rejects the null hypothesis under level α . That is, we report the value α^* such that the following equality holds:

$$\min_{\mathbf{u} \in \mathcal{U}} \max_{k \in \{1,2\}} \frac{t_k - \mu_{k,\mathbf{u}}}{\sigma_{k,\mathbf{u}}} = Q_{\rho_{\Gamma}^*, \alpha^*}.$$

It is clear that the adaptive test rejects the null hypothesis in a level-0.05 sensitivity

analysis conducted with sensitivity parameter Γ if and only if $\alpha^* \leq 0.05$. Like the worst-case p-value in a sensitivity analysis in the traditional sense, the value of α^* implies how strong the evidence against the null hypothesis obtained from the adaptive test is: a smaller α^* corresponds to a smaller chance that the null hypothesis holds in a sensitivity analysis.

Appendix H: More Details on Rosenbaum’s Adaptive Approach

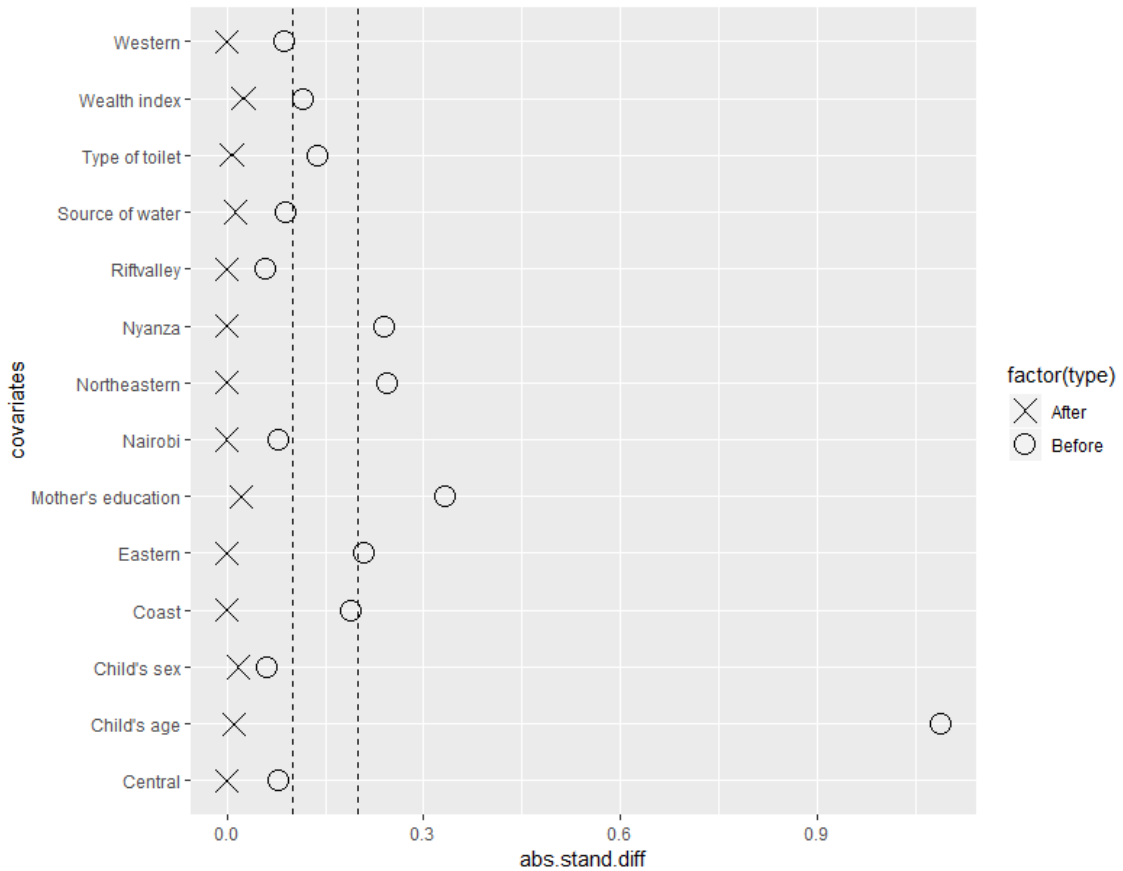
As mentioned in the main text, [Rosenbaum \(2012\)](#) designed an adaptive approach to combine two different test statistics in matched observational studies. The approach is data-driven and does not require dropping samples for design prior to statistical inference, and is designed to achieve the larger of the two design sensitivities of the two component tests. This traditional adaptive approach is designed for combining different tests within a large class of test statistics for pair matched samples, including any test statistics of the form $T = \sum_{i=1}^I \mathbb{1}(Y_i > 0) h_i$, where $Y_i = (Z_{i1} - Z_{i2})(R_{i1} - R_{i2})$ is the treated-minus-control difference in response for matched pair i and h_i is a function of $|Y_1|, \dots, |Y_I|$, in which case we can find a uniform upper bound test statistic \bar{T}_Γ under each sensitivity parameter Γ such that $\mathbb{P}(T \geq t \mid \mathcal{F}, \mathcal{Z}) \leq \mathbb{P}(\bar{T}_\Gamma \geq t \mid \mathcal{F}, \mathcal{Z})$ for any t . To combine two different test statistics, this traditional adaptive approach corrects for the correlation between the two test statistics by using the observation that the two upper bound statistics are asymptotically jointly normal under some regularity conditions; see Section 2 in [Rosenbaum \(2012\)](#) for details. [Rosenbaum \(2012\)](#) found that the cost for this correction is small compared with, for example, the Bonferroni adjustment since the two tests are typically highly correlated. However, as mentioned in the main text, Rosenbaum’s adaptive approach can only be applied to test statistics that are uniformly bounded by a known distribution, which typically requires either the

matching mechanism to be pair matching or the outcomes of interest to be binary, neither of which would hold for many widely used test statistics including the aberrant rank test, the Wilcoxon rank sum test, the Hodges-Lehmann aligned rank test or the Huber-Maritz m-tests ([Gastwirth et al., 2000](#); [Rosenbaum, 2002b, 2007](#)).

Appendix I: More Details on the Real Data Example

Here we give more details on the outcome of interest and the categorization of the observed covariates considered in the real data example described in Section 2.1.1.1 and Section 2.7. Height-for-age z-scores (the outcome variable) are expressed in units equal to one standard deviation of the reference population's distribution. As mentioned in Section 2.1.1.1, we match each treated individual with controls on the following seven observed covariates: mother's highest education level; geographic district; household wealth index in quantiles; household's main source of drinking water; household's toilet facilities; sex; and children's age in years. Mother's education is categorized as no education, primary, secondary and higher. Geographic district is coded as eight dummy variables with respect to eight districts in Kenya: Central, Coast, Eastern, Nairobi, Northeastern, Nyanza, Riftvalley and Western. We use a similar method to [Fink et al. \(2011\)](#) to code quality of source of drinking water and toilet facilities. As mentioned in the main text, to form the 150 matched sets, we applied optimal matching using rank-based Mahalanobis distance with a propensity score caliper ([Hansen and Klopfer, 2006](#)). Figure 2.1 shows the balance on the seven baseline covariates before and after matching evaluated by the standardized differences, defined as a weighted difference in means divided by the pooled standard deviation between the treated and control groups before matching (see Chapter 9 of [Rosenbaum \(2002b\)](#) for details). From Figure 2.1, we can see that the absolute standardized differences become much

Figure 2.1: Covariate imbalances before and after matching with three controls. The plot reports the absolute standardized differences before and after matching of each covariate. The two dotted vertical lines are 0.1 and 0.2 cut-offs.



smaller through matching and are all close to zero after matching.

Appendix J: More Discussions on the Aberrant Null

In this part, we give more discussions on the aberrant null and the aberrant rank test introduced in Section 2.4.1, especially, on how the individuals with normal (i.e., not aberrant) observed responses (including those with normal unobserved potential responses and those with aberrant unobserved potential responses) play a role in testing the aberrant null.

Unlike some traditional null hypotheses such as Fisher's sharp null, the aberrant

null only asserts that there is no treatment effect among individuals with aberrant potential responses either under treated or under control instead of among all individuals. However, this does not mean that the individuals with normal observed responses do not offer any information and should be discarded before testing. On the contrary, all individuals, no matter if their observed responses are normal or aberrant, can contribute information about the aberrant null, but in different ways. Under the aberrant null, for individuals with aberrant observed responses, their unobserved potential responses are the same as their observed responses and this hypothetical information is clearly significant for testing the aberrant null. For individuals with normal observed responses, although assuming the aberrant null is not sufficient for imputing the exact values of their unobserved potential responses, assuming the aberrant null at least implies that their unobserved potential responses are still normal no matter what. This information is useful and should be leveraged when testing the aberrant null. Conversely ignoring this information and discarding individuals with normal observed responses can potentially reduce the statistical power.

To better illustrate this last point about potential loss of statistical power if we discarded normal observed responses, consider testing the aberrant null $H_0^{\mathcal{A}}$ with $\mathcal{A} = [c, +\infty)$ for some threshold c

$$H_0^{\geq c} : r_{Tij} = r_{Cij}, \forall i, j, \text{ if either } r_{Tij} \geq c \text{ or } r_{Cij} \geq c,$$

with the aberrant rank test

$$T_{\text{abe}} = \sum_{i=1}^I \sum_{j=1}^m Z_{ij} q(R_{ij} | \mathbf{R}),$$

where $q(R_{ij} \mid \mathbf{R}) = \sum_{i'=1}^I \sum_{j'=1}^m \mathbb{1}(R_{ij} \geq R_{i'j'} \geq c)$ is the aberrant rank. As mentioned in Section 2.4.1, the aberrant rank of each individual is fixed under the aberrant null, and the aberrant ranks of individuals with normal observed responses all equal zero. Note that assigning aberrant rank zero to individuals with normal observed responses does not mean that we discarded these individuals in the test, as aberrant rank zero contains the important hypothetical information that, under the aberrant null, their potential responses under treated and under control are all normal. Also, we would reject the aberrant null if treated individuals significantly tend to have larger aberrant ranks than control individuals. More specifically, if there are two individuals A and B in the data, individual A has a larger aberrant rank than individual B if either one of the conditions hold:

- Case 1: both individuals A and B have aberrant observed responses, but individual A's observed response is more aberrant than that of individual B.
- Case 2: the observed response of individual A is aberrant but that of individual B is normal.

If we discarded all the normal observed responses in testing the aberrant null, we do not have the chance of leveraging the information from Case 2 listed above, which can potentially reduce the statistical power of our test. Therefore, all individuals should matter and should be involved when testing the aberrant null, regardless of if their observed responses were aberrant or not. For related discussions on the aberrant null and the aberrant rank test in completely randomized experiments, see [Rosenbaum and Silber \(2008\)](#).

3. A Model-Free and Finite-Population-Exact Framework for Randomized Experiments Subject to Outcome Misclassification via Integer Programming

This chapter is based on “Heng, S. and Shaw, P. A. (2021). A model-free and finite-population-exact framework for randomized experiments subject to outcome misclassification via integer programming. *arXiv:2201.03111*.”

3.1 Introduction

3.1.1 Outcome misclassification—a major source of bias in randomized experiments

For inferring causal effects of a treatment, conducting a randomized experiment (trial) is the gold standard. In a randomized experiment, treatments are randomly assigned according to some randomization design (i.e., assignment mechanism), and downstream statistical inference can be conducted based only on randomization (i.e., randomization-based inference) without imposing any super-population models on the subjects (e.g., assuming that subjects are i.i.d. realizations from

some data-generating models). See [Rosenbaum \(2002b, 2010\)](#), [Imbens and Rubin \(2015\)](#), and [Athey and Imbens \(2017\)](#) for detailed discussion. Although randomization can protect a causal conclusion from potential bias caused by confounders (measured or unmeasured) or model misspecification, it cannot remove potential bias resulting from measurement error in the outcome ([Arnold and Ercumen, 2016](#); Chapter 9; [Hernán and Robins, 2020](#)), referred to as *outcome misclassification* when the outcome is binary ([Carroll et al., 2006](#); [Buonaccorsi, 2010](#); [Yi, 2017](#)).

Outcome misclassification commonly exists in randomized experiments. For example, common binary outcomes in randomized clinical trials are cancer diagnosis (cancerous versus non-cancerous), blood test results (normal versus abnormal), antibody test results (positive versus negative), among many others. These binary clinical outcomes can be misclassified (misdiagnosed) due to one or several of the following factors: 1) technical limitations of the diagnosis equipment and methods ([Wittram et al., 2004](#)), 2) physician-side factors such as misinterpretations of the clinical data or patient's symptoms ([Walsh-Kelly et al., 1995](#)), and 3) patient-side factors such as patients' abnormal activities before the diagnosis and anxiety during the diagnosis ([Ogedegbe et al., 2008](#)). Another major type of outcome misclassification is reporting bias in self-reported binary outcomes, especially those concerning sensitive topics such as mental illness ([Knäuper and Wittchen, 1994](#)), sexuality variables ([Catania et al., 1986](#)), or unhappiness with a service or product ([Wood et al., 2008](#)).

In general, there are at least two ways that outcome misclassification, in particular systematic outcome misclassification (i.e., differential outcome misclassification by treatment status), can severely distort a causal conclusion drawn from a randomized experiment (trial). Firstly, in randomized trials without or with inade-

quate blinding, systematically larger effects on subjective outcomes (either subject-reported or investigator-assessed) often exist due to differential outcome misclassification from knowledge of treatment status, as found by a large systematic review of clinical trials (Wood et al., 2008). Such exaggerated reporting of health improvements documented in previous non-blinded or imperfectly blinded trials can be due to either placebo effects (i.e., treated subjects tend to perceive or overstate health improvements due to psychological factors) or courtesy bias (i.e., treated subjects tend to not fully state their unhappiness with the treatment as an attempt to be polite toward the investigators). However, perfect blinding is not practical in a wide range of randomized trials, and therefore bias due to outcome misclassification may be inevitable for many such studies (Arnold and Ercumen, 2016). Secondly, in some studies, the treatment can improve or interfere with the detection of the outcome. For example, some clinical trials have found that the treatments they investigated can change the volume of the study organ or the pattern of the study tumor, therefore diseases may be easier or harder to be detected among treated subjects than control subjects in such cases (Lucia et al., 2007; Redman et al., 2008). Therefore, even if blinding was effectively carried out in a randomized experiment (trial), potential bias due to systematic (differential) outcome misclassification may still exist.

3.1.1.1 Example: a puzzle from the Prostate Cancer Prevention Trial (PCPT)

Prostate cancer is one of the most common cancers in men. According to Siegel et al. (2020), prostate cancer is estimated to account for 21% of the new cancer cases diagnosed in men in the United States in 2020. Since the development of prostate cancer is a long-term process, many studies have focused on the prevention of prostate cancer. Among these studies, the Prostate Cancer Prevention Trial (PCPT)

(Thompson et al., 2003) is especially influential because it is “the first study to show that a drug can reduce a man’s chances of developing prostate cancer” (NIH, 2013). In the PCPT, within each of the 221 study sites, men 55 years of age or older with no evidence of prostate cancer are randomly assigned to finasteride (5 mg per day) (i.e., the treatment) or placebo (i.e., control). The primary outcome is whether the participant is diagnosed with prostate cancer during the seven-year follow-up period. Of the 9060 participants included in the final analysis, 803 of the 4368 in the finasteride group and 1147 of the 4692 in the placebo group were diagnosed with prostate cancer; see Table 1 in Thompson et al. (2003). Applying the Mantel-Haenszel test (randomization-based inference) to the 221 strata (study sites), the two-sided p-value under Fisher’s sharp null of no treatment effect is 4.66×10^{-13} . According to the two-sided 0.05 significance level prespecified by the PCPT (Thompson et al., 2003), a treatment effect of finasteride on the prevention of prostate cancer was detected.

The analysis reported by the PCPT also identified a controversial and seemingly contradictory phenomenon: taking finasteride may increase the risk of high-grade prostate cancer (i.e., tumor with Gleason score ≥ 7). Specifically, of the 9037 men (out of the total 9060 men in the final analysis) with available Gleason score, 280 of the 4358 in the finasteride group and 237 of the 4679 in the placebo group were diagnosed with high-grade prostate cancer. Applying the Mantel-Haenszel test, the two-sided p-value under Fisher’s sharp null is 6.79×10^{-3} , which is also statistically significant at the prespecified two-sided 0.05 level.

Can finasteride prevent prostate cancer but also promote high-grade prostate cancer? This seeming contradiction puzzled many researchers when the results were first published (Thompson et al., 2003). Several follow-up studies have pointed out

that this puzzling result could potentially be due to bias caused by misclassification of the prostate cancer status or severity (e.g., [Lucia et al., 2007](#); [Redman et al., 2008](#); [Shepherd et al., 2008](#)). So a natural question is: can we develop a model-free causal inference framework to help explain this puzzle from the perspective of outcome misclassification?

3.1.2 Our contributions

Although outcome misclassification commonly exists in randomized experiments, such as randomized clinical trials, and has the potential to severely distort a downstream causal conclusion, it is commonly ignored in practice. To the best of our knowledge, there is no established unified framework for randomized experiments subject to outcome misclassification without imposing any additional assumptions to a randomized experiment. Although there is extensive previous literature on addressing outcome misclassification in statistical inference, all of these previous approaches require assuming that subjects are realizations (typically i.i.d. realizations) from some super-population model (parametric, semiparametric, or nonparametric). For existing work, see [Quade et al. \(1980\)](#), [Magder and Hughes \(1997\)](#), [Lyles et al. \(2005\)](#), [Carroll et al. \(2006\)](#), [Küchenhoff et al. \(2006\)](#), [Shepherd et al. \(2008\)](#), [Buonaccorsi \(2010\)](#), [Gilbert et al. \(2016\)](#), [Yi \(2017\)](#), [Shu and Yi \(2019\)](#), [Beesley and Mukherjee \(2020\)](#), among many others. These model-based approaches have been successfully applied in various settings and greatly contributed to practical research. However, directly applying or adjusting these model-based approaches to randomized experiments may introduce unnecessary bias to a model-free randomization-based inference. On the one hand, for an approach based on some parametric or semiparametric model, the statistical inference may be biased due to model misspecification ([Magder and Hughes, 1997](#)).

On the other hand, even if an approach adopts a flexible nonparametric model, there may still exist potential bias due to the existence of unmeasured features that are involved in the outcome misclassification mechanism. For example, whether a patient's disease is correctly diagnosed or not highly depends on a doctor's expertise and experience on such disease, which may be difficult to measure or quantify (Weiss and Shanteau, 2003). Bias may also come from the violation of the i.i.d. assumption. Laboratories routinely calibrate their measuring instruments using the values of previously measured outcomes (Buonaccorsi, 2010). In such cases, whether some subject's outcome is correctly classified or not may depend on other subjects' outcomes, making the i.i.d. assumption unrealistic. Therefore, a model-free approach is needed for many randomized experiments.

In this paper, we develop a model-free and finite-population-exact framework for randomized experiments subject to outcome misclassification via integer programming. Our framework provides a unified approach to help address the following four common questions concerning the design, analysis, and validation of a randomized experiment:

- **Design Stage:** **Q1**—Given the planned randomization design, the planned sample size, and some assumed effect size, how accurate is accurate enough for the outcome measurement if we would like to ensure that the causal conclusion (e.g., rejecting a causal null hypothesis or not) based on the measured outcomes agrees with that based on the true outcomes?
- **Analysis Stage:** After forming a causal conclusion based on measured outcomes, two central questions are: **Q2**—How sensitive is this conclusion to outcome misclassification? **Q3**—Is this conclusion more sensitive to false positives vs. false negatives among the treated vs. control subjects?

- **Validation Stage: Q4**—For a validation substudy on outcome misclassification, how to choose which subjects to include to make the validation substudy more efficient?

A central concept of our framework is called *warning accuracy*, which is defined as the finite-population-exact threshold such that the causal conclusion based on the measured outcomes may differ from that based on the true outcomes if the accuracy of the measured outcomes did not surpass that threshold. We show how warning accuracy, the related *sensitive sets* and *sensitivity weights*, and its dual concept *design accuracy* can help investigate **Q1–4** without adding any additional assumptions to a randomized experiment. To handle the computational challenge encountered in computing warning accuracy for large-scale randomized experiments, called the “curse of symmetry,” we propose a computation strategy to adaptively reformulate a corresponding integer programming problem with respect to the randomization design. Our computation strategy leverages some intrinsic characteristics of various randomization designs and recent advances in erasing symmetry in integer programming (Fogarty et al., 2016, 2017), which can be of independent interest. Our framework covers both Fisher’s sharp null and Neyman’s weak null and works for a wide range of randomization designs. Our framework can also be applied to matched or stratified observational studies that adopt randomization-based inference. We illustrate our new framework through studying the well-known puzzle in the PCPT introduced in Section 3.1.1.1.

Model-based approaches in the previous literature on outcome misclassification can still be very useful (especially when we want to study questions beyond the scope of **Q1–Q4**) as long as the models and assumptions were appropriately imposed, and researchers can still do further model-based approaches after adopting

our framework. The central spirit of our framework is: *Before imposing any additional assumptions to a randomized experiment (meanwhile taking the risk of assumption violations or model misspecification) to conduct a model-assisted analysis for outcome misclassification, what useful information concerning outcome misclassification are we already be able to learn from the data?*

3.2 Review

3.2.1 Randomization-based inference with a binary outcome

Suppose that there are $I \geq 1$ strata (or blocks, or study centers), and there are n_i subjects in stratum i , $i = 1, \dots, I$, with $N = \sum_{i=1}^I n_i$ subjects in total. Let Z_{ij} be the treatment indicator of subject j in stratum i : $Z_{ij} = 1$ if subject j in stratum i received treatment and $Z_{ij} = 0$ otherwise. Suppose that in stratum i a fixed number m_i of subjects are designed to receive the treatment, and $n_i - m_i$ subjects receive control, i.e., $\sum_{j=1}^{n_i} Z_{ij} = m_i$ where $m_i \in \{1, \dots, n_i - 1\}$. Let $\mathbf{Z} = (Z_{11}, \dots, Z_{In_I})$ and $\mathcal{Z} = \{\mathbf{Z} \in \{0, 1\}^N : \sum_{j=1}^{n_i} Z_{ij} = m_i, i = 1, \dots, I\}$ denote all possible treatment assignments. In a randomized experiment, the treatment assignments are random within each stratum, i.e.,

$$P(\mathbf{Z} = \mathbf{z} \mid \mathcal{Z}) = \prod_{i=1}^I \binom{n_i}{m_i}^{-1}, \quad \forall \mathbf{z} \in \mathcal{Z}. \quad (3.1)$$

In an observational study, the randomization assumption (3.1) is often assumed after adjusting for confounders using matching or stratification (Rosenbaum, 2002b). The above set-up covers a wide range of randomized experiments and observational studies. When $I = 1$, the study is called a completely randomized experiment. For $I \geq 2$, the study is a general stratified/blocked randomized experiment

or observational study. When $n_i = 2$ for all i , the study is a paired randomized experiment or pair-matched observational study. When $m_i = 1$ and $n_i = n \geq 3$ for all i , the study is a randomized experiment or observational study with multiple controls. When $m_i = 1$ for all i while n_i may vary with i , the study is a randomized experiment or observational study with variable controls. When $\min\{m_i, n_i - m_i\} = 1$ for all i , the study is a finely stratified experiment (Fogarty, 2018) or an observational study with full matching (Rosenbaum, 1991; Hansen, 2004). See Imbens and Rubin (2015) and Rosenbaum (2002b, 2010) for detailed introductions and definitions of various types of randomized experiments and observational studies.

Let $Y_{ij} \in \{0, 1\}$ be the true outcome of subject j in stratum i and $\mathbf{Y} = (Y_{11}, \dots, Y_{I n_I})$. Following the potential outcomes framework (Neyman, 1923; Rubin, 1974), let $Y_{ij}(1)$ and $Y_{ij}(0)$ be the potential (true) outcome under treated and that under control of subject j in stratum i respectively, therefore $Y_{ij} = Y_{ij}(1)Z_{ij} + Y_{ij}(0)(1 - Z_{ij})$. In randomization-based inference for randomized experiments or observational studies, potential outcomes are *fixed values* and the only probability distribution that enters statistical inference is the randomization assumption (3.1) (Rosenbaum, 2002b; Imbens and Rubin, 2015; Athey and Imbens, 2017). Fisher's sharp null hypothesis of no treatment effect (Fisher, 1935) asserts that $H_0^{\text{sharp}} : Y_{ij}(1) = Y_{ij}(0), \forall i, j$. The commonly used test statistics for testing H_0^{sharp} in randomization-based inference with binary outcome are Fisher's exact test (when $I = 1$) and the Mantel-Haenszel test (when $I \geq 1$), which is defined as $T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y}) = \sum_{i=1}^I \sum_{j=1}^{n_i} Z_{ij} Y_{ij}$ (Mantel and Haenszel, 1959). The Mantel-Haenszel test reduces to Fisher's exact test when $I = 1$ and reduces to McNemar's test when $n_i = 2$ for all i (Cox and Snell, 2018). People commonly use the following finite-population

central limit theorem (Hájek, 1960; Rosenbaum, 2002b; Li and Ding, 2017) to test H_0^{sharp} : $\frac{T_{\text{M-H}} - E(T_{\text{M-H}})}{\sqrt{\text{Var}(T_{\text{M-H}})}} \xrightarrow{\mathcal{L}} N(0, 1)$, where $E(T_{\text{M-H}}) = \sum_{i=1}^I \left(\frac{m_i}{n_i} \sum_{j=1}^{n_i} Y_{ij}\right)$ and $\text{Var}(T_{\text{M-H}}) = \sum_{i=1}^I \frac{m_i(\sum_{j=1}^{n_i} Y_{ij})(n_i - \sum_{j=1}^{n_i} Y_{ij})(n_i - m_i)}{n_i^2(n_i - 1)}$. In a two-sided level- α testing procedure, people reject H_0^{sharp} if and only if $\frac{\{T_{\text{M-H}} - E(T_{\text{M-H}})\}^2}{\text{Var}(T_{\text{M-H}})} > \chi_{1,1-\alpha}^2$, where $\chi_{1,1-\alpha}^2$ is $1 - \alpha$ quantile of chi-squared distribution with one degree of freedom.

Another extensively considered null hypothesis is Neyman's weak null hypothesis of no average treatment effect $H_0^{\text{weak}} : \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (Y_{ij}(1) - Y_{ij}(0)) = 0$ (Neyman, 1923), which can be rewritten as $H_0^{\text{weak}} : \sum_{i=1}^I \frac{n_i}{N} \tau_i = 0$, where $\tau_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}(1) - \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}(0)$. A commonly used test statistic for testing H_0^{weak} is the Neyman estimator (i.e., the difference-in-means estimator) (Neyman, 1923): $T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y}) = \sum_{i=1}^I \frac{n_i}{N} \hat{\tau}_i$, where $\hat{\tau}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} Z_{ij} Y_{ij} - \frac{1}{n_i - m_i} \sum_{j=1}^{n_i} (1 - Z_{ij}) Y_{ij}$. Under H_0^{weak} , we have $E(T_{\text{Neyman}}) = 0$ and $\text{Var}(T_{\text{Neyman}}) = \sum_{i=1}^I \left(\frac{n_i}{N}\right)^2 \text{Var}(\hat{\tau}_i)$, where $\text{Var}(\hat{\tau}_i) = \frac{S_{T,i}^2}{m_i} + \frac{S_{C,i}^2}{n_i - m_i} - \frac{S_i^2}{n_i}$, with $S_{T,i}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij}(1) - \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}(1))^2$, $S_{C,i}^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij}(0) - \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}(0))^2$, and $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij}(1) - Y_{ij}(0) - \tau_i)^2$. People commonly use the following finite-population central limit theorem (Hájek, 1960; Imbens and Rubin, 2015; Li and Ding, 2017) to test H_0^{weak} : $\frac{T_{\text{Neyman}}}{\sqrt{\text{Var}(T_{\text{Neyman}})}} \xrightarrow{\mathcal{L}} N(0, 1)$. Since S_i^2 involves τ_i which cannot be identified from the observed data, people commonly adopt the following conservative estimator $\widehat{\text{Var}}(T_{\text{Neyman}})$ for $\text{Var}(T_{\text{Neyman}})$ (Neyman, 1923): $\widehat{\text{Var}}(T_{\text{Neyman}}) = \sum_{i=1}^I \left(\frac{n_i}{N}\right)^2 \cdot \left(\frac{\widehat{S}_{T,i}^2}{m_i} + \frac{\widehat{S}_{C,i}^2}{n_i - m_i}\right)$, where $\widehat{S}_{T,i}^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} Z_{ij} (Y_{ij}^* - \frac{1}{m_i} \sum_{j=1}^{m_i} Z_{ij} Y_{ij}^*)^2$ and $\widehat{S}_{C,i}^2 = \frac{1}{n_i - m_i - 1} \sum_{j=1}^{n_i} (1 - Z_{ij}) \{Y_{ij}^* - \frac{1}{n_i - m_i} \sum_{j=1}^{n_i} (1 - Z_{ij}) Y_{ij}^*\}^2$. For example, in a two-sided level- α testing procedure, people reject H_0^{weak} if and only if $\frac{T_{\text{Neyman}}^2}{\widehat{\text{Var}}(T_{\text{Neyman}})} > \chi_{1,1-\alpha}^2$. See Imbens and Rubin (2015) for more details.

3.2.2 Some basic concepts about outcome misclassification

As discussed in Section 3.1.1, in practice the measured outcomes $\mathbf{Y}^* = (Y_{11}^*, \dots, Y_{In_I}^*)$ may be subject to misclassification (i.e., $\mathbf{Y}^* \neq \mathbf{Y}$), and a causal conclusion based on \mathbf{Y}^* may differ from that based on the true outcomes \mathbf{Y} . In outcome misclassification and binary classification literature, the measured outcome Y_{ij}^* is said to be a *true positive* if $(Y_{ij}^*, Y_{ij}) = (1, 1)$, a *false positive* if $(Y_{ij}^*, Y_{ij}) = (1, 0)$, a *true negative* if $(Y_{ij}^*, Y_{ij}) = (0, 0)$, or a *false negative* if $(Y_{ij}^*, Y_{ij}) = (0, 1)$. Let TP, FP, TN and FN denote the total number of subjects that lie in the above four categories respectively. One of the most fundamental and widely used measures of the precision of \mathbf{Y}^* is *accuracy*, which is defined as the proportion of correct classification among \mathbf{Y}^* under the true outcomes \mathbf{Y} (denoted as $\mathcal{A}(\mathbf{Y}^* | \mathbf{Y})$):

$$\mathcal{A}(\mathbf{Y}^* | \mathbf{Y}) = \frac{TP + TN}{TP + FP + TN + FN} = \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} \mathbf{1}(Y_{ij}^* = Y_{ij})}{N}. \quad (3.2)$$

Note that accuracy $\mathcal{A}(\mathbf{Y}^* | \mathbf{Y})$ is a model-free and finite-population-exact concept and is therefore especially compatible with randomization-based inference framework which does not require assuming any super-population models on the study subjects.

3.3 A Model-Free and Finite-Population-Exact Framework

In this section, we introduce several model-free and finite-population-exact quantities to help address the four common questions **Q1–4** listed in Section 3.1.2. In Section 3.3.1, we give the motivations, definitions, and explanations of those quan-

ties. In Section 3.3.2, we conclude how to use those quantities to address Q1–4.

3.3.1 Warning accuracy, sensitivity weights, and design accuracy

After obtaining a causal conclusion (i.e., rejecting causal null hypothesis or not) based on the measured outcomes, a natural question is: how many misclassified outcomes are needed at least to overturn that causal conclusion? We call such number as *minimal alteration number*. Or equivalently, what is the threshold such that the causal conclusion based on the measured outcomes may differ from that based on the true outcomes if the accuracy of the measured outcomes did not surpass that threshold? We call this threshold the *warning accuracy*, for which a rigorous definition is given below.

Definition 1 (Warning Accuracy). Let $\mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y}^*)$ and $\mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y})$ denote the level- α hypothesis testing result (rejecting the null hypothesis or not) based on the measured outcomes \mathbf{Y}^* and that based on the true outcomes \mathbf{Y} respectively, and $\mathcal{A}(\mathbf{Y}^* | \mathbf{Y})$ the accuracy of \mathbf{Y}^* under \mathbf{Y} . The warning accuracy given level α , the treatment indicators \mathbf{Z} , and the measured outcomes \mathbf{Y}^* , is defined as

$$\mathcal{WA} = \max_{\mathbf{Y}: \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y}^*) \neq \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y})} \mathcal{A}(\mathbf{Y}^* | \mathbf{Y}). \quad (3.3)$$

Then the corresponding minimal alteration number is $(1 - \mathcal{WA}) \times N$.

We here give some remarks on the warning accuracy \mathcal{WA} defined in Definition 1. First, it immediately follows from Definition 1 that if the (unknown) actual outcome accuracy $\mathcal{A}(\mathbf{Y}^* | \mathbf{Y}) > \mathcal{WA}$, we have $\mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y}^*) = \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y})$ while if $\mathcal{A}(\mathbf{Y}^* | \mathbf{Y}) \leq \mathcal{WA}$, it may happen that $\mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y}^*) \neq \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y})$. Therefore, other things being equal, a smaller warning accuracy \mathcal{WA} indicates less sensitivity to

outcome misclassification. Second, the warning accuracy is model-free and finite-population-exact as it does not require assuming that subjects are realizations from some super-population model.

We now use two simple examples to illustrate why reporting the warning accuracy, as a complement of the p -value based on measured outcomes, can greatly benefit a randomized experiment (or an observational study using randomization-based inference) subject to outcome misclassification.

Example 1. *Consider a complete randomized experiment with one treated subject and 1000 controls. Without loss of generality, suppose that the treatment indicators $\mathbf{Z} = (1, 0, \dots, 0)$. Assume that the corresponding measured outcomes $\mathbf{Y}^* = (1, 0, \dots, 0)$. Then the p -value under Fisher's sharp null based on \mathbf{Y}^* is $1/1001 < 0.001$, which would be considered as very strong evidence of treatment effect for many scientific journals. However, if the true outcomes $\mathbf{Y} = (0, 0, \dots, 0)$, the true p -value will be one (no evidence), even if \mathbf{Y}^* and \mathbf{Y} only differ by one case of misclassification. In this case, the warning accuracy is $1000/1001$, implying high sensitivity to outcome misclassification, even if the p -value based on \mathbf{Y}^* is very statistically significant.*

Example 2. *We consider the following two stratified randomized experiments with equal total sample size ($=17$) and the same prespecified level of 0.05 (one-sided). It is easy to check that study 1 has a smaller one-sided p -value (under Fisher's sharp null) but larger warning accuracy while study 2 has a larger one-sided p -value but smaller warning accuracy.*

| | | | | | |
|-----------|-----------------|---------------|-------------------|---------|-------|
| Study 1 | Stratum 1 | Stratum 2 | Stratum 3 | p-value | WA |
| Z | (1 0 1) | (0 0 1 0 0 0) | (0 0 0 1 0 0 0 0) | 1/144 | 16/17 |
| Y* | (1 0 1) | (0 0 1 0 0 0) | (0 0 0 1 0 0 0 0) | | |
| Study 2 | Stratum 1 | Stratum 2 | Stratum 3 | p-value | WA |
| Z | (1 0 0 0 0 0 0) | (1 1 0 1 0) | (0 0 1 1 0) | 1/100 | 15/17 |
| Y* | (0 0 0 0 0 0 0) | (1 1 0 1 0) | (0 0 1 1 0) | | |

In summary, Example 1 shows that even a causal conclusion with a very small p-value can be very sensitive to outcome misclassification (exhibits high warning accuracy). Example 2 implies that a causal conclusion with a smaller p-value is not necessarily less sensitive to outcome misclassification than that with a larger p-value. Therefore, when outcome misclassification is a concern, reporting the warning accuracy can provide useful information about sensitivity to outcome misclassification. Such information cannot be covered by p-value based on measured outcomes and does not require any additional assumptions.

Reporting warning accuracy alone has two limitations. First, warning accuracy is a worst-case scenario sensitivity analysis for outcome misclassification. While the worst-case scenario is universally valid, prior information or expert knowledge may be able to shed light as to whether the warning accuracy is overly conservative. Second, warning accuracy itself does not provide any information about if a causal conclusion is in general more sensitive to a false positive versus a false negative in the treated versus the control group. We propose another concept called *sensitivity weights* to overcome these two limitations, which is built on the definition of warning accuracy and the following concept called a *sensitive set*. A sensitive set refers to a minimal (in terms of cardinality) set of subjects such that if

the outcomes of those subjects were misclassified, the causal conclusion based on the measured outcomes will be overturned. Therefore, it is a collection of subjects whose outcomes being misclassified or not are particularly influential to the causal conclusion. A formal definition of a sensitive set is given below.

Definition 2 (Sensitive Set). *Under the setting of Definition 1, let*

$$\tilde{\mathbf{Y}} = (\tilde{Y}_{11}, \dots, \tilde{Y}_{I_1 I_1}) \in \underset{\mathbf{Y}: \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y}^*) \neq \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y})}{\operatorname{argmax}} \mathcal{A}(\mathbf{Y}^* | \mathbf{Y})$$

be an optimal solution to the optimization problem (3.3) associated with the definition of warning accuracy. Then $\mathcal{S} = \{ij : \tilde{Y}_{ij} \neq Y_{ij}^\}$ is called a sensitive set.*

In practice, there may exist more than one sensitive set as the solution to the optimization problem involved in Definition 1 may not be unique. Therefore, a sensitive set itself may not be an intrinsic quantify. However, as we will show in Section 3.4, for the most widely used tests such as the Mantel-Haenszel test and the Neyman estimator, these different sensitive sets can, in general, be transformed to each other by some composition of within-strata and/or between-strata permutations, and share the same quantities called *sensitivity wights*—the key to answer **Q3** and **Q4**.

Definition 3 (Sensitivity Weights). *Under the setting of Definition 1, let \mathcal{S} be a sensitive set. Then there is a set of sensitivity weights defined as the following 2×2 table:*

| <i>Sensitivity Weights</i> | <i>False Positives</i> | <i>False Negatives</i> |
|----------------------------|------------------------|------------------------|
| <i>Treated</i> | W_T^{FP} | W_T^{FN} |
| <i>Control</i> | W_C^{FP} | W_C^{FN} |

where W_T^{FP} , W_T^{FN} , W_C^{FP} , and W_C^{FN} denote the sensitivity weight of false positives in the

treated group, that of false negatives in the treated group, that of false positives in the control group, and that of false negatives in the control group respectively, defined as

$$W_T^{FP} = \frac{|\{ij : Z_{ij} = 1, Y_{ij}^* = 1, \tilde{Y}_{ij} = 0\}|}{|\mathcal{S}|}, \quad W_T^{FN} = \frac{|\{ij : Z_{ij} = 1, Y_{ij}^* = 0, \tilde{Y}_{ij} = 1\}|}{|\mathcal{S}|},$$

$$W_C^{FP} = \frac{|\{ij : Z_{ij} = 0, Y_{ij}^* = 1, \tilde{Y}_{ij} = 0\}|}{|\mathcal{S}|}, \quad W_C^{FN} = \frac{|\{ij : Z_{ij} = 0, Y_{ij}^* = 0, \tilde{Y}_{ij} = 1\}|}{|\mathcal{S}|}.$$

The rationale of sensitivity weights is straightforward: since the causal conclusion is particularly sensitive to potential cases of outcome misclassification among the subjects in a sensitive set, then the proportion of each of the four types of outcome misclassification (i.e., false positives/negatives in the treated/control group) within a sensitive set offers a sensible quantification about if the causal conclusion is in particular sensitive to certain type of outcome misclassification. For example, in Example 1, it is clear that the sensitivity weight of a false positive in the treated group is the dominant term according to Definition 3, suggesting that the causal conclusion is especially sensitive to a false positive in the treated group and therefore should be given priority when conducting a validation study.

Not only can our framework benefit the analysis and validation of a randomized experiment or an observational study based on randomization-based inference, but also it can provide a benchmark for outcome measurement accuracy at the design stage of such studies. Specifically, by introducing a dual concept of warning accuracy called *design accuracy*, under each assumed effect size, we can calculate the expected threshold such that the causal conclusion based on the measured outcomes is guaranteed to agree with that based on the true outcomes as long as the outcome measurement accuracy surpasses that threshold. We call such procedure in a design stage the *accuracy calculation*.

Definition 4 (Design Accuracy). Let $\mathbf{Y} \sim (p_0, p_1)$ denote that $Y_{ij} \mid Z_{ij} = 0 \sim \text{Bern}(p_0)$ and $Y_{ij} \mid Z_{ij} = 1 \sim \text{Bern}(p_1)$. The design accuracy given level α , the treatment indicators \mathbf{Z} , and the assumed effect size (p_0, p_1) , is defined as

$$\mathcal{DA} = E_{\mathbf{Y} \sim (p_0, p_1)} \left\{ \max_{\mathbf{Y}^*: \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y}^*) \neq \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y})} \mathcal{A}(\mathbf{Y}^* \mid \mathbf{Y}) \right\}.$$

As indicated in Definition 4, design accuracy can be seen as a dual concept of warning accuracy. For warning accuracy, measured outcomes are given and we use an optimization solver to manipulate the unknown true outcomes to conduct a worst-case scenario sensitivity analysis. In contrast, for design accuracy, we generate each set of true outcomes according to some assumed effect size and manipulate measured outcomes for each generated set of true outcomes to conduct a worst-case scenario accuracy calculation. Ideally, we would like outcome measurement to be as accurate as possible. However, there is typically a trade-off between budget and outcome accuracy – the more accurate the outcome measurement, the more budget and resources we may need to achieve that accuracy (Lubovsky et al., 2005). Design accuracy provides a sensible benchmark for outcome accuracy without modeling the outcome misclassification mechanism.

3.3.2 Strengthening a randomized experiment with warning accuracy, sensitivity weights, and design accuracy

Putting the concepts developed in Section 3.3.1 together, we show how to use warning accuracy, sensitivity weights, and design accuracy to strengthen the design, analysis, and validation of a randomized experiment subject to outcome misclassification.

Design Stage: To address **Q1**, design accuracy can serve as a worst-case scenario benchmark for outcome accuracy. Specifically, after fixing the sample size and randomization design, researchers can specify various effect sizes (p_0, p_1) and generate 1000 simulated sets of true outcomes \mathbf{Y} according to $Y_{ij} \mid Z_{ij} = 0 \sim \text{Bern}(p_0)$ and $Y_{ij} \mid Z_{ij} = 1 \sim \text{Bern}(p_1)$. Then researchers compute the 1000 corresponding optimal values of the following optimization problem: $\max_{\mathbf{Y}^*: \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y}^*) \neq \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y})} \mathcal{A}(\mathbf{Y}^* \mid \mathbf{Y})$, and take their average as the approximate design accuracy defined in Definition 4.

Analysis Stage: To address **Q2**, after reporting the p-value based on the measured outcomes, researchers can then report the warning accuracy under the prespecified level. A low warning accuracy implies a causal conclusion's high insensitivity (robustness) to outcome misclassification. If the warning accuracy is relatively high instead, researchers should report the sensitivity weights and check if the dominant term among the four types of outcome misclassification (false positives/negatives in the treated/control group) agrees with that based on prior information and/or expert knowledge or not (we will illustrate such procedure in detail in Section 3.5). If this is the case, the causal conclusion drawn from the measured outcomes may be misleading even if the unknown actual outcome accuracy equals or is close to that high warning accuracy. Otherwise, we would expect the actual accuracy needed to overturn the causal conclusion based on the measured outcomes to be lower, or even much lower, than that high warning accuracy. To address **Q3**, we report the four sensitivity weights and observe if there is any dominant term among the four sensitivity weights. Such a term would be a strong sign that the causal conclusion based on the measured outcomes is particularly sensitive to that type of outcome misclassification.

Validation Stage: If the warning accuracy is relatively high and the dominant term among the four sensitivity weights is expected in practice, then researchers may want to select a subset of study subjects for outcome validation. When choosing a validation subset, it makes sense to give priorities to (i) the type of outcome misclassification with the dominant sensitivity weight and (ii) the subjects that belong to a sensitive set. This offers a sensible answer to **Q4**.

3.4 Computing Warning Accuracy and Related Quantities

3.4.1 The original problem formulation and the “curse of symmetry”

When the sample size N is large, it is typically infeasible to calculate warning accuracy by hand as in Section 3.3.1. A general strategy for tackling a computationally extensive problem involving integers, such as calculating warning accuracy (3.3), is to appropriately formulate the problem into an integer program and apply a state-of-the-art optimization solver. We first consider testing H_0^{sharp} with the routinely used Mantel-Haenszel test, which reduces to Fisher’s exact test when $I = 1$. If H_0^{sharp} was rejected by the Mantel Haenszel test at level α , i.e., if $\frac{[T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y}^*) - E\{T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y}^*)\}]^2}{\text{Var}\{T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y}^*)\}} > \chi_{1,1-\alpha}^2$ by Definition 1, the warning accuracy is the optimal value of the following integer quadratically constrained linear program:

$$\underset{\mathbf{Y} \in \{0,1\}^N}{\text{maximize}} \quad \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} Y_{ij}^* Y_{ij} + \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*)(1 - Y_{ij}) \quad (\text{P0})$$














$$\text{subject to} \quad [T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y}) - E\{T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y})\}]^2 - \chi_{1,1-\alpha}^2 \cdot \text{Var}\{T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y})\} \leq 0.$$

The objective function of (P0) comes from Definition 1 and a simple observation that $\sum_{i=1}^I \sum_{j=1}^{n_i} \mathbf{1}(Y_{ij}^* = Y_{ij}) = \sum_{i=1}^I \sum_{j=1}^{n_i} Y_{ij}^* Y_{ij} + \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*)(1 - Y_{ij})$, which is a linear function in \mathbf{Y} given \mathbf{Y}^* . The inequality constraint of (P0) comes from the fact that $\mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y}) \neq \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y}^*) = \text{“Reject”}$ if and only if $\frac{[T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y}) - E\{T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y})\}]^2}{\text{Var}\{T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y})\}} \leq \chi_{1,1-\alpha}^2$, which can be rewritten as a quadratic constraint in \mathbf{Y} as in (P0). Instead if H_0^{sharp} fails to be rejected by the Mantel-Haenszel test based on \mathbf{Y}^* , to compute the warning accuracy, we just need to solve a simple variant of (P0) with replacing the “ ≤ 0 ” with the “ ≥ 0 ” in the constraint. Typically, a sensitivity analysis (e.g., calculating warning accuracy) makes more sense if a null hypothesis was rejected (i.e., some treatment effect was detected) by a primary analysis (Rosenbaum, 2002b), so in the rest of this paper we focus on solving (P0). For calculating the warning accuracy under Neyman’s weak null, we just need to replace the inequality constraint in (P0) with $\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y})\}^2 - \chi_{1,1-\alpha}^2 \cdot \widehat{\text{Var}}\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y})\} \leq 0$ (or ≥ 0), which is still a quadratic constraint in \mathbf{Y} . See Appendix A.3 for more details.

The original problem formulation (P0) seems straightforward and natural. However, this seemingly reasonable integer program formulation can easily make the computation infeasible because of the so-called “curse of symmetry” (Margot, 2010), as will be explained further below. For clarity of the notation, without loss of generality, in Section 3.4.1 we temporarily re-organize subjects within the same stratum such that for each stratum i , we have $Z_{ij_1} \leq Z_{ij_2}$ and $Y_{ij_1}^* \leq Y_{ij_2}^*$ as long as $j_1 \leq j_2$. Following the typical notation in group theory (Scott, 2012), we let (ij, ij') denote the permutation of the index set $\mathcal{I} = \{11, 12, \dots, In_I\}$ such that index ij exchanges with index ij' while all other indexes remain the same. Define the following permutation group G_{within} over \mathcal{I} : $G_{\text{within}} = \{\text{All possible compositions of any } (ij, ij') \text{ s.t. } Y_{ij}^* = Y_{ij'}^* \text{ and } Z_{ij} = Z_{ij'}\}$. Let $g\mathbf{Y}$ denote that the permu-

tation g acts on the indexes in \mathbf{Y} : $g\mathbf{Y} = g(Y_{11}, \dots, Y_{In_I}) = (Y_{g(11)}, \dots, Y_{g(In_I)})$. It is clear that for any $g \in G_{\text{within}}$, we have $\mathcal{A}(\mathbf{Y}^* \mid g\mathbf{Y}) = \mathcal{A}(\mathbf{Y}^* \mid \mathbf{Y})$ and $\mathcal{D}_\alpha(\mathbf{Z}, g\mathbf{Y}) = \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y})$, i.e., the integer program (P0) is invariant under any permutation $g \in G_{\text{within}}$. We call this property *within-strata symmetry*. Then for each stratum i , let $\Lambda_i = (\Lambda_i^{00}, \Lambda_i^{01}, \Lambda_i^{10}, \Lambda_i^{11}) = (\sum_{j=1}^{n_i} (1 - Z_{ij})(1 - Y_{ij}^*), \sum_{j=1}^{n_i} (1 - Z_{ij})Y_{ij}^*, \sum_{j=1}^{n_i} Z_{ij}(1 - Y_{ij}^*), \sum_{j=1}^{n_i} Z_{ij}Y_{ij}^*)$ denote the measured 2×2 table of stratum i . Let S denote the number of unique 2×2 tables among $\{\Lambda_i, i = 1, \dots, I\}$, and let P_s denote the number of strata with the measured 2×2 tables equalling the s -th unique table $\Lambda_{[s]} = (\Lambda_{[s]}^{00}, \Lambda_{[s]}^{01}, \Lambda_{[s]}^{10}, \Lambda_{[s]}^{11}), s = 1, \dots, S$, therefore $\sum_{s=1}^S P_s = I$. For two strata i and i' with $n_i = n_{i'}$, we let $(i, i') = (i1, i'1) \dots (in_i, i'n_{i'})$ denote the permutation of \mathcal{I} such that stratum i 's indexes element-wisely exchange with stratum i' 's indexes while all other indexes remain the same. Define the following permutation group $G_{\text{between}} = \{\text{All possible compositions of any } (i, i') \text{ s.t. } \Lambda_i = \Lambda_{i'}\}$ over \mathcal{I} . It is clear that for any $g \in G_{\text{between}}$, we also have $\mathcal{A}(\mathbf{Y}^* \mid g\mathbf{Y}) = \mathcal{A}(\mathbf{Y}^* \mid \mathbf{Y})$ and $\mathcal{D}_\alpha(\mathbf{Z}, g\mathbf{Y}) = \mathcal{D}_\alpha(\mathbf{Z}, \mathbf{Y})$, i.e., the integer program (P0) is invariant under any permutation $g \in G_{\text{between}}$. We call such property as *between-strata symmetry*. To illustrate, we classify all the study subjects into four types based on the treatment indicator Z and measured outcome Y^* as in Table 3.1 (a). Then an illustration of the two types of symmetry, within-strata symmetry and between-strata symmetry, is given in Table 3.1 (b).

Table 3.1: Illustration of the two types of symmetry: between-strata symmetry (e.g., stratum 1 and stratum 2) and within-strata symmetry (e.g., subject 1 and subject 2 in stratum 3).

| | | | | | | | |
|---------|---|---|--|-----------|--|---|---|
| | $Y^* = 1$ | $Y^* = 0$ | | | Subject 1 | Subject 2 | Subject 3 |
| $Z = 1$ |  |  | | Stratum 1 |  |  |  |
| $Z = 0$ |  |  | | Stratum 2 |  |  |  |
| | | | | Stratum 3 |  |  |  |

(a) Four types of study subjects

(b) Two types of symmetry

Putting the above discussions together, define the permutation group $G = \{\text{All possible compositions of elements from } G_{\text{within}} \text{ and } G_{\text{between}}\}$. For any $g \in G$, the integer program (P0) is invariant under permutation g . Note that

$$|G| = |G_{\text{between}}| \times |G_{\text{within}}| = \prod_{s=1}^S P_s! \times \prod_{i=1}^I \Lambda_i^{00}! \Lambda_i^{01}! \Lambda_i^{10}! \Lambda_i^{11}!,$$

which is an extremely large number if N is large, indicating an extremely high degree of symmetry of program (P0) and resulting in computational infeasibility of solving (P0) due to the so-called “curse of symmetry,” which refers to a general fact that an integer program is typically computationally infeasible if its variables can be permuted in many ways (e.g., $|G|$ is large) without changing the structure of the problem (Margot, 2010).

In addition to the computational challenge, another implication of the above arguments is that, as mentioned in Section 3.3.1, a sensitive set (i.e., an optimal solution to the integer program (P0)) itself is not an intrinsic concept because for a given sensitive set, its transformation under some between-strata permutations (as in G_{between}) and/or within-strata permutations (as in G_{within}) is still a sensitive set.

However, a simple but important observation is that these different sensitive sets have the same sensitivity weights. As mentioned in Section 3.3.1, this motivates us to introduce an intrinsic concept associated with a sensitive set – sensitivity weights, which is invariant under any permutations in G_{between} and G_{within} and therefore provides an intrinsic quantification of a causal conclusion’s relative sensitivity to the four different types of outcome misclassification.

3.4.2 Two types of randomization designs and an adaptive reformulation strategy

In Section 3.4.2, we propose a general strategy to solve the “curse of symmetry” encountered when calculating the warning accuracy according to the original problem formulation (P0). The core idea of our strategy is to reformulate the integer program (P0) with respect to an intrinsic characteristic of various randomization designs—whether within-strata symmetry dominates between-strata symmetry for that randomization design or vice versa. Specifically, we classify many commonly used randomization designs into the following two types:

- **Type I randomization designs:** those in which within-strata symmetry dominates between-strata symmetry (i.e., $|G_{\text{within}}| \gg |G_{\text{between}}|$). This class of randomization designs include some commonly used randomization designs (including randomized experiments and observational studies adopting randomization-based inference) such as completely randomized experiments (for which $|G_{\text{between}}| = 1$) and stratified/blocked randomized experiments or observational studies with most strata/blocks being large.
- **Type II randomization designs:** those in which between-strata symmetry

dominates within-strata symmetry (i.e., $|G_{\text{between}}| \gg |G_{\text{within}}|$). This class of randomization designs include some widely used randomization designs such as paired randomized experiments or pair-matched observational studies (for which $|G_{\text{within}}| = 1$), randomized experiments or observational studies with multiple controls, randomized experiments or observational studies with variable controls, finely stratified experiments or observational studies with full matching, and stratified/blocked randomized experiments or observational studies with most strata/blocks being small.

3.4.2.1 Type I: within-strata symmetry dominates between-strata symmetry

We first show how to reformulate the integer program (P0) for type I randomization designs. By independence of treatment assignments between strata, the definition of $\mathcal{A}(\mathbf{Y} \mid \mathbf{Y}^*)$ and that of the Mantel-Haenszel test, the accuracy (the objective function of (P0)) can be determined by the I 2×2 tables $\{\sum_{j=1}^{n_i} \mathbf{1}(Y_{ij}^* = b, Y_{ij} = c) : b, c \in \{0, 1\}\}$ ($i = 1, \dots, I$) and the constraint's feasibility (i.e., the Mantel-Haenszel test fails to reject H_0^{sharp} based on \mathbf{Y}) can be determined by the I 2×2 tables $\{\sum_{j=1}^{n_i} \mathbf{1}(Z_{ij} = a, Y_{ij} = c) : b, c \in \{0, 1\}\}$ ($i = 1, \dots, I$). Therefore, the integer program (P0) can be determined by the I $2 \times 2 \times 2$ tables $\{\sum_{j=1}^{n_i} \mathbf{1}(Z_{ij} = a, Y_{ij}^* = b, Y_{ij} = c) : a, b, c \in \{0, 1\}\}$ ($i = 1, \dots, I$). Because \mathbf{Z} and \mathbf{Y}^* have been observed, for each stratum i , the i -th $2 \times 2 \times 2$ table only has the following four degrees of freedom: $Y_i^{00} = \sum_{j=1}^{n_i} \mathbf{1}(Z_{ij} = 0, Y_{ij}^* = 0, Y_{ij} = 1)$, $Y_i^{01} = \sum_{j=1}^{n_i} \mathbf{1}(Z_{ij} = 0, Y_{ij}^* = 1, Y_{ij} = 1)$, $Y_i^{10} = \sum_{j=1}^{n_i} \mathbf{1}(Z_{ij} = 1, Y_{ij}^* = 0, Y_{ij} = 1)$, and $Y_i^{11} = \sum_{j=1}^{n_i} \mathbf{1}(Z_{ij} = 1, Y_{ij}^* = 1, Y_{ij} = 1)$. Let $\mathbf{Y} = (Y_1^{00}, Y_1^{01}, Y_1^{10}, Y_1^{11}, \dots, Y_I^{00}, Y_I^{01}, Y_I^{10}, Y_I^{11}) \in \mathbb{Z}^{4I}$ and $\check{Y}_i = Y_i^{00} + Y_i^{01} + Y_i^{10} + Y_i^{11}$, then (P0) can be reformulated as the following integer

quadratic program (P1):

$$\begin{aligned}
\max_{\mathbf{Y} \in \mathbb{Z}^{4I}} \quad & \frac{1}{N} \sum_{i=1}^I (Y_i^{01} + Y_i^{11} - Y_i^{00} - Y_i^{10}) + \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*) \quad (\text{P1}) \\
\text{s.t.} \quad & \left\{ \sum_{i=1}^I (Y_i^{10} + Y_i^{11}) - \sum_{i=1}^I \frac{m_i}{n_i} \check{Y}_i \right\}^2 - \chi_{1,1-\alpha}^2 \cdot \sum_{i=1}^I \frac{m_i \check{Y}_i (n_i - \check{Y}_i) (n_i - m_i)}{n_i^2 (n_i - 1)} \leq 0, \\
& 0 \leq Y_i^{00} \leq \sum_{j=1}^{n_i} (1 - Z_{ij}) (1 - Y_{ij}^*), \quad \forall i \\
& 0 \leq Y_i^{01} \leq \sum_{j=1}^{n_i} (1 - Z_{ij}) Y_{ij}^*, \quad \forall i \\
& 0 \leq Y_i^{10} \leq \sum_{j=1}^{n_i} Z_{ij} (1 - Y_{ij}^*), \quad \forall i \\
& 0 \leq Y_i^{11} \leq \sum_{j=1}^{n_i} Z_{ij} Y_{ij}^*. \quad \forall i
\end{aligned}$$

Note that in (P1), in addition to rewriting the objective function and constraint of (P0) in terms of new variables \mathbf{Y} , we also add the bounding constraints for \mathbf{Y} . It is clear that there is no within-strata symmetry anymore in (P1), which was previously a major source of symmetry in the original formulation (P0) for type I randomization designs. The above argument also works for the case of testing Neyman's weak null H_0^{weak} with type I randomization designs; see Appendix B.3 for details. In Appendix C, we showed how to calculate sensitivity weights and a collection of sensitive sets after solving (P1).

3.4.2.2 Type II: between-strata symmetry dominates within-strata symmetry

Our strategy of reformulating the integer program (P0) for type II randomization designs is inspired by both the observations discussed in Section 3.4.2.1 and an idea from [Fogarty et al. \(2016\)](#). Specifically, in [Fogarty et al. \(2016\)](#), to erase sym-

metry in finding the worst-case variance estimator for the average treatment effect with binary outcomes in an observational study with full matching, instead of manipulating the potential outcomes based on the treatment indicators and observed outcomes, the authors proposed to list all possible $2 \times 2 \times 2$ tables (in terms of treatment indicators, observed outcomes, and potential outcomes) based on observed data and manipulate the number of each possible $2 \times 2 \times 2$ table that enter into randomization inference. Although in this paper we are studying a totally different problem, we can combine this general philosophy and the arguments in Section 3.4.2.1 to reformulate the integer program (P0) for type II randomization designs.

A high-level summary of the idea is: for type II randomization designs, there are many “words” (strata), but not many “vocabularies” (all possible $2 \times 2 \times 2$ tables in terms of binary treatment indicators, binary measured outcomes, and binary true outcomes, given the observed data). Therefore, instead of directly manipulating the potential $2 \times 2 \times 2$ table for each stratum as in (P1), we first create a “dictionary” that lists all possible unique $2 \times 2 \times 2$ tables based on measured 2×2 tables (in terms of treatment indicators and measured outcomes), and then manipulate the total number of strata that take certain $2 \times 2 \times 2$ table. We call this strategy as the “dictionary method.” Specifically, for the s -th unique 2×2 table $\Lambda_{[s]} = (\Lambda_{[s]}^{00}, \Lambda_{[s]}^{01}, \Lambda_{[s]}^{10}, \Lambda_{[s]}^{11}), s = 1, \dots, S$, let \tilde{n}_s and \tilde{m}_s denote its number of total study subjects and its number of treated subjects. Moreover, for the s -th unique 2×2 table $\Lambda_{[s]}$, there are $\tilde{N}_s = (\Lambda_{[s]}^{00} + 1)(\Lambda_{[s]}^{01} + 1)(\Lambda_{[s]}^{10} + 1)(\Lambda_{[s]}^{11} + 1)$ possible unique $2 \times 2 \times 2$ tables $\{\sum_{j=1}^{\tilde{N}_s} \mathbf{1}(Z_{ij} = a, Y_{ij}^* = b, Y_{ij} = c) : a, b, c \in \{0, 1\}\}$ ($s = 1, \dots, S$). Let d_{sp} be the number of the p -th unique $2 \times 2 \times 2$ table Δ_{sp} for the s -th unique table $\Lambda_{[s]}$, $s = 1, \dots, S$ and $p = 1, \dots, \tilde{N}_s$. Since \mathbf{Z} and \mathbf{Y}^*

are given, each $2 \times 2 \times 2$ table Δ_{sp} can be uniquely determined by four numbers $\Delta_{sp}^{00} = \sum_{j=1}^{\tilde{n}_s} \mathbf{1}(Z_{ij} = 0, Y_{ij}^* = 0, Y_{ij} = 1)$, $\Delta_{sp}^{01} = \sum_{j=1}^{\tilde{n}_s} \mathbf{1}(Z_{ij} = 0, Y_{ij}^* = 1, Y_{ij} = 1)$, $\Delta_{sp}^{10} = \sum_{j=1}^{\tilde{n}_s} \mathbf{1}(Z_{ij} = 1, Y_{ij}^* = 0, Y_{ij} = 1)$, and $\Delta_{sp}^{11} = \sum_{j=1}^{\tilde{n}_s} \mathbf{1}(Z_{ij} = 1, Y_{ij}^* = 1, Y_{ij} = 1)$. Let $\check{\Delta}_{sp} = \Delta_{sp}^{00} + \Delta_{sp}^{01} + \Delta_{sp}^{10} + \Delta_{sp}^{11}$, then we can reformulate the integer program (P0) as the following integer quadratic program (P2):

$$\begin{aligned}
\max_{d_{sp} \in \mathbb{Z}} \quad & \frac{1}{N} \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} (\Delta_{sp}^{01} + \Delta_{sp}^{11} - \Delta_{sp}^{00} - \Delta_{sp}^{10}) + \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*) \quad (\text{P2}) \\
\text{s.t.} \quad & \left\{ \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} (\Delta_{sp}^{10} + \Delta_{sp}^{11}) - \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} \left(\frac{\tilde{m}_s}{\tilde{n}_s} \cdot \check{\Delta}_{sp} \right) \right\}^2 \\
& - \chi_{1,1-\alpha}^2 \cdot \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} \frac{\tilde{m}_s \check{\Delta}_{sp} (\tilde{n}_s - \check{\Delta}_{sp}) (\tilde{n}_s - \tilde{m}_s)}{\tilde{n}_s^2 (\tilde{n}_s - 1)} \leq 0, \\
& \sum_{p=1}^{\tilde{N}_s} d_{sp} = P_s, \quad \forall s \\
& d_{sp} \geq 0. \quad \forall s, p
\end{aligned}$$

Note that in (P2), $(\Delta_{sp}^{00}, \Delta_{sp}^{01}, \Delta_{sp}^{10}, \Delta_{sp}^{11})$ are fixed numbers (i.e., ‘‘vocabularies’’ listed in a ‘‘dictionary’’) and the decision variables are d_{sp} , i.e., the total number of strata that take the unique $2 \times 2 \times 2$ table $(\Delta_{sp}^{00}, \Delta_{sp}^{01}, \Delta_{sp}^{10}, \Delta_{sp}^{11})$. It is clear that there is neither within-strata symmetry nor between-strata symmetry in (P2) and therefore surpasses the formulation (P1) in terms of erasing symmetry. However, the dimension of the decision variables in (P2) is $\sum_{s=1}^S \tilde{N}_s = \sum_{s=1}^S (\Lambda_{[s]}^{00} + 1)(\Lambda_{[s]}^{01} + 1)(\Lambda_{[s]}^{10} + 1)(\Lambda_{[s]}^{11} + 1)$, which is typically very high for type I randomization designs, but is typically not high for type II randomization. Therefore, for type II randomization designs, it is appropriate to reformulate (P0) as (P2). While for type I randomization designs for which within-strata symmetry is the main concern, it is more appropriate to reformulate (P0) as (P1).

The above argument also works for the case of testing H_0^{weak} with type II randomization designs; see Appendix B.4 for details. A method of calculating sensitivity weights and a collection of sensitive sets after solving (P2) is given in Appendix C.

3.4.3 Simulation studies

We conduct simulations to study the computational efficiency of the adaptive reformulation strategy proposed in Section 3.4.2. We also gain some insights on how warning accuracy and sensitivity weights vary with the effect size of measured outcomes and sample size. Note that the data generating processes described below are only intended for automatically generating data sets for simulations as our framework works for any given data sets and does not require assuming any data generating models. We investigate both type I and type II randomization designs through considering the following two simulation scenarios:

- **Simulation Scenario 1 (for Type I randomization designs):** We consider a stratified randomized experiment (or a stratified observational study adopting randomization-based inference) with $I = 40$ or 200 strata. We let $\mathcal{U}(A)$ denote uniform distribution over the set A . In each independent simulation run, for each $i = 1, \dots, I$ we randomly draw m_i from $\mathcal{U}(\{10, 11, \dots, 40\})$ and then randomly draw $n_i - m_i$ from $\mathcal{U}(\{10, 11, \dots, 40\})$. The expected total number of study subjects $E(N) = \{E(m_i) + E(n_i - m_i)\} \cdot I = 50I = 2000$ or $10,000$.
- **Simulation Scenario 2 (for Type II randomization designs):** We consider a finely stratified randomized experiment (or a matched observational study with full matching) with $I = 400$ or 2000 strata. In each independent simulation run, for each $i = 1, \dots, I$ we randomly generate m_i and $n_i - m_i$

based on the following procedure: we first set $\min\{m_i, n_i - m_i\} = 1$ and draw $\max\{m_i, n_i - m_i\}$ from $\mathcal{U}(\{1, 2, \dots, 7\})$ and then randomly assign $\min\{m_i, n_i - m_i\}$ and $\max\{m_i, n_i - m_i\}$ to m_i and $n_i - m_i$. Then we have $E(N) = \{E(m_i) + E(n_i - m_i)\} \cdot I = 5I = 2000$ or $10,000$.

For both Simulation Scenarios 1 and 2, in each independent simulation run we generate \mathbf{Z} and \mathbf{Y}^* according to the following procedure: 1) Given the generated n_i and m_i , the treatments are randomly assigned within each stratum, i.e., the randomization assumption (3.1) holds. 2) We then independently generate a measured outcome Y_{ij}^* for each study subject ij according to: $Y_{ij}^* \sim \text{Bernoulli}(p_1)$ if $Z_{ij} = 1$ and $Y_{ij}^* \sim \text{Bernoulli}(p_0)$ if $Z_{ij} = 0$. We here consider testing Fisher's sharp null H_0^{sharp} . The parallel simulation studies for Neyman's weak null H_0^{weak} are reported in Appendix D in the supplementary materials. After conducting 1000 independent simulation runs for each of the different prespecified sets of $(E(N), p_0, p_1)$ (18 sets in total) under Simulation Scenarios 1 and 2, we report the corresponding average computation time, average warning accuracy and average sensitivity weights in Table 3.2. Here are some further details about the specific procedure of obtaining the results in Table 3.2: (i) As mentioned in Section 3.4.1, conducting a sensitivity analysis or a validation study typically makes more sense when a treatment effect was detected in a primary analysis (Rosenbaum, 2002b) based on measured outcomes \mathbf{Y}^* . Therefore, we exclude few simulation runs (i.e., generated data sets) in which the null hypothesis failed to be rejected based on the generated measured outcomes (107 out of 36,000 runs). (ii) For the remaining 35,893 simulation runs, since the worst-case complexity of integer programming is \mathcal{NP} -hard, to avoid our simulation study failing to be finished in a tolerable amount of time, we force a simulation run to stop if it exceeds 100 seconds, report the total num-

ber of such cases, and exclude such cases when calculating the average computation time, warning accuracy, and sensitivity weights. However, such potentially computationally infeasible cases (computation time exceeds 100 seconds) are very rare (5 out of 35,893 runs) and our framework is computationally efficient in most cases. (iii) The computation was done by the optimization solver Gurobi (version 9.1) (Gurobi Optimization, LLC, 2022) and a laptop computer with a 1.6 GHz Dual-Core Intel Core i5 processor and 4 GB 1600 MHz DDR3 memory. See Table 3.2 for the detailed simulation results.

Table 3.2: The average computation time (in seconds), warning accuracy \mathcal{WA} and sensitivity weights $(W_T^{FP}, W_T^{FN}, W_C^{FP}, W_C^{FN})$ of different sets of $(E(N), p_0, p_1)$ for Simulation Scenarios 1 and 2 (for type I and type II randomization designs respectively).

| Type I Randomization Designs (Simulation Scenario 1) | | | | | | | | | | | | |
|---|---------------|----------------|------------|------------|------------|------------|-----------------|----------------|------------|------------|------------|------------|
| $p_0 = 0.3$ | $E(N) = 2000$ | | | | | | $E(N) = 10,000$ | | | | | |
| | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.4$ | 0.15 s | 0.98 | 0.33 | 0.00 | 0.00 | 0.67 | 3.67 s | 0.97 | 0.35 | 0.00 | 0.00 | 0.65 |
| $p_1 = 0.6$ | 0.17 s | 0.91 | 0.46 | 0.00 | 0.00 | 0.54 | 3.73 s | 0.90 | 0.46 | 0.00 | 0.00 | 0.54 |
| $p_1 = 0.8$ | 0.18 s | 0.83 | 0.54 | 0.00 | 0.00 | 0.46 | 3.79 s | 0.82 | 0.54 | 0.00 | 0.00 | 0.46 |
| $p_0 = 0.6$ | $E(N) = 2000$ | | | | | | $E(N) = 10,000$ | | | | | |
| | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.7$ | 0.15 s | 0.98 | 0.68 | 0.00 | 0.00 | 0.32 | 3.66 s | 0.97 | 0.66 | 0.00 | 0.00 | 0.34 |
| $p_1 = 0.8$ | 0.16 s | 0.95 | 0.72 | 0.00 | 0.00 | 0.28 | 3.73 s | 0.94 | 0.69 | 0.00 | 0.00 | 0.31 |
| $p_1 = 0.9$ | 0.17 s | 0.91 | 0.75 | 0.00 | 0.00 | 0.25 | 3.71 s | 0.90 | 0.72 | 0.00 | 0.00 | 0.28 |
| $p_0 = 0.9$ | $E(N) = 2000$ | | | | | | $E(N) = 10,000$ | | | | | |
| | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.2$ | 0.21 s | 0.74 | 0.00 | 0.46 | 0.54 | 0.00 | 4.15 s | 0.74 | 0.00 | 0.47 | 0.53 | 0.00 |
| $p_1 = 0.4$ | 0.19 s | 0.83 | 0.00 | 0.37 | 0.63 | 0.00 | 3.79 s | 0.82 | 0.00 | 0.38 | 0.62 | 0.00 |
| $p_1 = 0.6$ | 0.17 s | 0.91 | 0.00 | 0.25 | 0.75 | 0.00 | 3.70 s | 0.90 | 0.00 | 0.28 | 0.72 | 0.00 |
| Type II Randomization Designs (Simulation Scenario 2) | | | | | | | | | | | | |
| $p_0 = 0.3$ | $E(N) = 2000$ | | | | | | $E(N) = 10,000$ | | | | | |
| | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.4$ | 2.75 s | 0.99 | 0.13 | 0.00 | 0.00 | 0.87 | 6.33 s | 0.99 | 0.27 | 0.00 | 0.00 | 0.73 |

| | | | | | | | | | | | | |
|-------------|--------|--------------------------|------------|------------|------------|------------|-----------------|--------------------------|------------|------------|------------|------------|
| $p_1 = 0.6$ | 2.68 s | 0.96 | 0.44 | 0.00 | 0.00 | 0.56 | 8.00 s | 0.95 | 0.44 | 0.00 | 0.00 | 0.56 |
| $p_1 = 0.8$ | 1.95 s | 0.92 | 0.54 | 0.00 | 0.00 | 0.46 | 8.68 s | 0.91 | 0.54 | 0.00 | 0.00 | 0.46 |
| | | $E(N) = 2000$ | | | | | $E(N) = 10,000$ | | | | | |
| $p_0 = 0.6$ | | | | | | | | | | | | |
| | Time | $\mathcal{W}\mathcal{A}$ | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | $\mathcal{W}\mathcal{A}$ | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.7$ | 2.80 s | 0.99 | 0.85 | 0.00 | 0.00 | 0.15 | 6.43 s | 0.99 | 0.73 | 0.00 | 0.00 | 0.27 |
| $p_1 = 0.8$ | 2.22 s | 0.97 | 0.76 | 0.00 | 0.00 | 0.24 | 6.19 s | 0.97 | 0.72 | 0.00 | 0.00 | 0.28 |
| $p_1 = 0.9$ | 1.45 s | 0.96 | 0.75 | 0.00 | 0.00 | 0.25 | 5.40 s | 0.95 | 0.75 | 0.00 | 0.00 | 0.25 |
| | | $E(N) = 2000$ | | | | | $E(N) = 10,000$ | | | | | |
| $p_0 = 0.9$ | | | | | | | | | | | | |
| | Time | $\mathcal{W}\mathcal{A}$ | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | $\mathcal{W}\mathcal{A}$ | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.2$ | 1.09 s | 0.88 | 0.00 | 0.47 | 0.53 | 0.00 | 6.86 s | 0.87 | 0.00 | 0.47 | 0.53 | 0.00 |
| $p_1 = 0.4$ | 1.45 s | 0.92 | 0.00 | 0.38 | 0.62 | 0.00 | 7.33 s | 0.91 | 0.00 | 0.39 | 0.61 | 0.00 |
| $p_1 = 0.6$ | 1.44 s | 0.96 | 0.00 | 0.25 | 0.75 | 0.00 | 5.41 s | 0.95 | 0.00 | 0.26 | 0.74 | 0.00 |

From the simulation results shown in Table 3.2, we can find that: 1) In most simulation runs, our computation strategy for calculating warning accuracy and sensitivity weights is very efficient with the average computation time being a few seconds, even for large data sets (e.g., $E(N) = 10,000$). 2) Other things being equal, warning accuracy decreases as measured effect size (i.e., difference in p_0 and p_1) increases, which agrees with the fact that detection of a treatment effect is less sensitive to outcome misclassification if the measured effect size is larger. 3) Specific values of sensitivity weights should depend on the specific randomization design, sample size, and (p_0, p_1) , but we can still observe some general patterns from the simulations. First, when a positive treatment effect was detected (i.e., $p_1 > p_0$ and H_0^{sharp} was rejected), only W_T^{FP} and W_C^{FN} can be non-zero (in other words, W_T^{FN} and W_C^{FP} must be zero). This agrees with the simple fact that there are only two types of outcome misclassification that can overturn a detected positive treatment effect: false positives among the treated group and false negatives among

the control group. Similarly, when a negative treatment effect was detected (i.e., $p_1 < p_0$ and H_0^{sharp} was rejected), only W_T^{FN} and W_C^{FP} can be non-zero (in other words, W_T^{FP} and W_C^{FN} must be zero). Second, among the two non-zero sensitivity weights, which one dominates the other depends on the specific randomization design and (p_0, p_1) . For example, for both Simulation Scenarios 1 and 2, when $p_0 = 0.3$ and $p_1 = 0.4$, W_C^{FN} dominates W_T^{FP} , while when $p_0 = 0.6$ and $p_1 = 0.9$, W_T^{FP} dominates W_C^{FN} instead. In some cases, the two non-zero sensitivity weights are comparable; see $p_0 = 0.9$ and $p_1 = 0.2$ for Simulation Scenarios 1 and 2.

3.5 Real Data Application: Understanding the Puzzle in the PCPT

We now apply our newly developed framework to understand the puzzle in the PCPT described in Section 3.1.1.1. Following [Thompson et al. \(2003\)](#), the prespecified alpha level is 0.05. We apply the efficient computation strategy developed in Section 3.4 to calculate warning accuracy (equivalently, minimal alteration number) and sensitivity weights for the two binary outcomes of interest in the PCPT: prostate cancer indicator (cancer versus no cancer) and high-grade prostate cancer indicator (high-grade prostate cancer versus no cancer or low-grade cancer) at 7 years. The results are reported in Table 3.3 below.

Table 3.3: The p-values, warning accuracy and sensitivity weights for the two binary outcomes of interest in the PCPT under Fisher’s sharp null hypothesis of no treatment effect and alpha level 0.05.

| Outcome (Sample Size N) | Prostate Cancer ($N = 9060$) | High-Grade Prostate Cancer ($N = 9037$) |
|----------------------------|------------------------------------|--|
| Relative Risk | 0.75 (Protective Factor) | 1.27 (Risk Factor) |
| p-value | 4.66×10^{-13} | 6.79×10^{-3} |
| Reject H_0 or Not | Yes | Yes |
| Causal Conclusion | Prevents Prostate Cancer | Promotes High-Grade Cancer |
| Warning Accuracy | 98.37% | 99.88% |
| Minimal Alteration # | 147 | 11 |
| Sensitivity Weights | False Positives False Negatives | False Positives False Negatives |
| Finasteride | 0 132/147 | 2/11 0 |
| Placebo | 15/147 0 | 0 9/11 |

We now give an interpretation of the results in Table 3.3. First, according to the reported values of warning accuracy and minimal alteration number, in a worst-case scenario sensitivity analysis, the causal conclusion concerning the prevention effect of Finasteride on prostate cancer is less sensitive to outcome misclassification than the causal conclusion concerning the promotion effect on high-grade prostate cancer (the two values of warning accuracy differ by 1.5%). A 1.5% difference in warning accuracy is nontrivial as the sample size is large (over 9000 study subjects) and it corresponds to a difference of 136 in the minimal alteration number. In other words, to alter the causal conclusion concerning the prevention effect, it requires $147/11 \approx 13.4$ times more cases of outcome misclassification than that required by the causal conclusion concerning the promotion effect.

Second, we leverage the reported sensitivity weights and related prior information

and expert knowledge to further investigate sensitivity to outcome misclassification for each of the two causal conclusions. From the reported sensitivity weights in Table 3.3, we learn that for both of the causal conclusions, the major concern is false negatives: the dominant term among the sensitivity weights for the prevention effect is false negatives among the Finasteride (treated) group and that for the promotion effect is false negatives among the placebo (control) group. For each of the two causal conclusions, is it plausible that the dominant term among the sensitivity weights is the dominant term among the four actual numbers of outcome misclassification cases? We now use related prior information and expert knowledge to shed light on this issue for the two outcomes of interest. We define the following notation: N_T —number of treated subjects; N_C —number of control subjects; $p_{T,1}$ (or $p_{T,0}$)—proportion of positive (or negative) true outcomes among the treated subjects; $p_{C,1}$ (or $p_{C,0}$)—proportion of positive (or negative) true outcomes among the control subjects; $\pi_{T,1|0}$ (or $\pi_{T,0|1}$)—proportion of false positives (or false negatives) among the treated subjects with true outcomes being negative (or positive); $\pi_{C,1|0}$ (or $\pi_{C,0|1}$)—proportion of false positives (or false negatives) among the control subjects with true outcomes being negative (or positive). Then the total number of each of the four types of outcome misclassification (false positives/negatives among the treated/control group) can be decomposed into the product of these three terms, as shown in Table 3.4.

Table 3.4: Decomposition of the total number of outcome misclassification cases for each of the four types of outcome misclassification.

| Misclassification Cases | False Positives | False Negatives |
|-------------------------|---------------------------------------|---------------------------------------|
| Finasteride (Treated) | $N_T \cdot p_{T,0} \cdot \pi_{T,1 0}$ | $N_T \cdot p_{T,1} \cdot \pi_{T,0 1}$ |
| Placebo (Control) | $N_C \cdot p_{C,0} \cdot \pi_{C,1 0}$ | $N_C \cdot p_{C,1} \cdot \pi_{C,0 1}$ |

For PCPT, we have the following related prior information or expert knowledge: (i) We have $N_T \approx N_C$ by design. (ii) Even if there are misclassified outcomes, it may still be sensible to get some sense of the values of $(p_{T,0}, p_{T,1}, p_{C,0}, p_{C,1})$ using measured outcomes. Based on measured outcomes, for the prostate cancer (all grades) outcome, $p_{T,0} \approx 82\%$, $p_{T,1} \approx 18\%$, $p_{C,0} \approx 76\%$, $p_{C,1} \approx 24\%$. For the high-grade prostate cancer outcome, $p_{T,0} \approx 94\%$, $p_{T,1} \approx 6\%$, $p_{C,0} \approx 95\%$, $p_{C,1} \approx 5\%$. Here “ \approx ” means a very rough estimation based on measured outcomes. (iii) Finasteride substantially improves detection of prostate cancer (all-grades). According to [NIH \(2013\)](#): “Finasteride has several effects on the prostate that allow better detection of prostate cancers. The drug shrinks the prostate, reducing its size and volume and increasing the chance that a biopsy will find existing cancers.” Meanwhile, Finasteride also greatly improves accuracy in prostate cancer grading at biopsy ([Redman et al., 2008](#)). Therefore, we expect $\pi_{C,0|1} \gg \pi_{T,0|1}$ for both outcomes.

Therefore, according to (i), (ii), and (iii), we expect that: 1) For the prostate cancer outcome, false negatives among the treated (Finasteride) group (i.e., the dominant term among the four sensitivity weights) should not be the actual dominant term among the four types of outcome misclassification, as at least we expect that the number of false negatives among the treated should not dominate the number of false negatives among controls. Consequently, we expect that for the prostate cancer outcome, the actual accuracy needed to overturn the causal conclusion about the prevention effect should be lower than the warning accuracy of 98.37%. 2) For the high-grade prostate cancer outcome, false negatives among the control (placebo) group (i.e., the dominant term among the four sensitivity weights) could be the actual dominant term among the four types of outcome misclassification, as we at least expect that the number of false negatives among controls should

dominate the number of false negatives among the treated. Although whether the number of false negatives among controls dominates the number of false positives among the treated/control group needs further investigation and information, unlike the prostate cancer outcome (prevention effect), for the high-grade prostate cancer outcome we currently do not have evidence that the actual accuracy needed to overturn the causal conclusion about the promotion effect should be lower than the reported warning accuracy of 99.88%.

Putting all the arguments together, based on the experimental data, the information provided using our new approach, and related expert knowledge, we have evidence that the causal conclusion that Finasteride prevents prostate cancer is more reliable and the causal conclusion that Finasteride promotes high-grade prostate cancer is less convincing and may be due to systematic (differential) outcome misclassification. Our finding supports some previous arguments proposed by domain scientists and biostatisticians (e.g., [Lucia et al., 2007](#); [Redman et al., 2008](#); [Shepherd et al., 2008](#)).

3.6 Summary

In this paper, we proposed a model-free and finite-population-exact framework for answering some widely concerned questions in a randomized experiment subject to outcome misclassification, covering the design stage, analysis stage, and validation stage. The strength of our framework is that it does not require any additional assumptions and meanwhile can provide some useful information to help researchers analyze a randomized experiment in a more comprehensive way. Our problem formulation strategy (i.e., adaptive integer program formulation with respect to the randomization design) for handling the “curse of symmetry” en-

countered in our framework could shed light on other computationally intensive problems in causal inference with binary outcomes.

As emphasized in Section 3.1.2, model-based approaches to outcome misclassification are very useful and necessary when researchers plan to use additional side information and domain knowledge to investigate the questions beyond those covered in our framework (i.e., **Q1-Q4** summarized in Section 3.1.2). The motivation of our framework is: what useful information concerning outcome misclassification can we learn from the experimental data before making any modeling or distributional assumptions? Such information, although may not be able to address all the issues concerning outcome misclassification, is immune to any assumption violations and model misspecification and is always trustworthy.

3.7 Appendices

Appendix A: Review and Some Preliminary Results

A.1: General form of an integer quadratically constrained linear program

In integer programming literature, the general standard form of an integer quadratically constrained linear program ([Lee and Leyffer, 2011](#); [Burer and Sax-](#)

ena, 2012; Conforti et al., 2014) is as below:

$$\begin{aligned}
& \underset{\mathbf{x}}{\text{maximize}} && \mathbf{q}^T \mathbf{x} + c \quad (\text{linear objective function}) \\
& \text{subject to} && \mathbf{x}^T \mathbf{Q}_k \mathbf{x} + \mathbf{q}_k^T \mathbf{x} \leq b_k, \forall k \quad (\text{quadratic constraints}) \\
& && \mathbf{A} \mathbf{x} \leq \mathbf{b}, \quad (\text{linear constraints}) \\
& && \mathbf{l} \leq \mathbf{x} \leq \mathbf{u}, \quad (\text{box constraints}) \\
& && \text{Some or all elements of } \mathbf{x} \text{ are integers.} \quad (\text{integrality constraints})
\end{aligned}$$

A.2: Some preliminary calculations concerning computing warning accuracy with testing Fisher's sharp null

In this section, we give some preliminary results for writing the integer programs (P0), (P1) and (P2) as the standard form given in Appendix A.1. Specifically, we would like to rewrite every quadratic term appearing in some quadratic constraint in each integer program considered in this paper as the standard form $\mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{q}^T \mathbf{x}$, which is required by optimization solvers such as Gurobi (Gurobi Optimization, LLC, 2022). For the quadratic constraint in the integer program (P0), we have

$$\begin{aligned}
& [T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y}) - E\{T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y})\}]^2 - \chi_{1,1-\alpha}^2 \cdot \text{Var}\{T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y})\} \\
& = \left\{ \sum_{i=1}^I \sum_{j=1}^{n_i} Z_{ij} Y_{ij} - \sum_{i=1}^I \left(\frac{m_i}{n_i} \sum_{j=1}^{n_i} Y_{ij} \right) \right\}^2 \\
& \quad - \chi_{1,1-\alpha}^2 \cdot \sum_{i=1}^I \frac{m_i (\sum_{j=1}^{n_i} Y_{ij}) (n_i - \sum_{j=1}^{n_i} Y_{ij}) (n_i - m_i)}{n_i^2 (n_i - 1)} \\
& = \sum_{i=1}^I \sum_{j=1}^{n_i} \left\{ \left(Z_{ij} - \frac{m_i}{n_i} \right)^2 + \chi_{1,1-\alpha}^2 \cdot \frac{m_i (n_i - m_i)}{n_i^2 (n_i - 1)} \right\} Y_{ij}^2 \\
& \quad + \sum_{i=1}^I \sum_{j \neq j'} \left\{ \left(Z_{ij} - \frac{m_i}{n_i} \right) \left(Z_{ij'} - \frac{m_i}{n_i} \right) + \chi_{1,1-\alpha}^2 \cdot \frac{m_i (n_i - m_i)}{n_i^2 (n_i - 1)} \right\} Y_{ij} Y_{ij'}
\end{aligned}$$

$$\begin{aligned}
& + \sum_{i \neq i'} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} \left(Z_{ij} - \frac{m_i}{n_i} \right) \left(Z_{i'j'} - \frac{m_{i'}}{n_{i'}} \right) Y_{ij} Y_{i'j'} \\
& - \sum_{i=1}^I \sum_{j=1}^{n_i} \chi_{1,1-\alpha}^2 \cdot \frac{m_i n_i (n_i - m_i)}{n_i^2 (n_i - 1)} Y_{ij}.
\end{aligned}$$

The above equation can be rewritten as

$$\begin{aligned}
& [T_{M-H}(\mathbf{Z}, \mathbf{Y}) - E\{T_{M-H}(\mathbf{Z}, \mathbf{Y})\}]^2 - \chi_{1,1-\alpha}^2 \cdot \text{Var}\{T_{M-H}(\mathbf{Z}, \mathbf{Y})\} \\
& = \left\{ \sum_{i=1}^I \sum_{j=1}^{n_i} Z_{ij} Y_{ij} - \sum_{i=1}^I \left(\frac{m_i}{n_i} \sum_{j=1}^{n_i} Y_{ij} \right) \right\}^2 \\
& \quad - \chi_{1,1-\alpha}^2 \cdot \sum_{i=1}^I \frac{m_i (\sum_{j=1}^{n_i} Y_{ij}) (n_i - \sum_{j=1}^{n_i} Y_{ij}) (n_i - m_i)}{n_i^2 (n_i - 1)} \\
& = \left[\sum_{i=1}^I (Y_i^{10} + Y_i^{11}) - \sum_{i=1}^I \left\{ \frac{m_i}{n_i} (Y_i^{00} + Y_i^{01} + Y_i^{10} + Y_i^{11}) \right\} \right]^2 \\
& \quad - \chi_{1,1-\alpha}^2 \cdot \sum_{i=1}^I \frac{m_i (Y_i^{00} + Y_i^{01} + Y_i^{10} + Y_i^{11}) (n_i - Y_i^{00} - Y_i^{01} - Y_i^{10} - Y_i^{11}) (n_i - m_i)}{n_i^2 (n_i - 1)} \\
& = \sum_{i=1}^I \left[\left\{ \frac{m_i^2}{n_i^2} + \chi_{1,1-\alpha}^2 \cdot \frac{m_i (n_i - m_i)}{n_i^2 (n_i - 1)} \right\} (Y_i^{00})^2 \right. \\
& \quad + \left\{ \frac{m_i^2}{n_i^2} + \chi_{1,1-\alpha}^2 \cdot \frac{m_i (n_i - m_i)}{n_i^2 (n_i - 1)} \right\} (Y_i^{01})^2 \\
& \quad + \left\{ \left(1 - \frac{m_i}{n_i} \right)^2 + \chi_{1,1-\alpha}^2 \cdot \frac{m_i (n_i - m_i)}{n_i^2 (n_i - 1)} \right\} (Y_i^{10})^2 \\
& \quad + \left\{ \left(1 - \frac{m_i}{n_i} \right)^2 + \chi_{1,1-\alpha}^2 \cdot \frac{m_i (n_i - m_i)}{n_i^2 (n_i - 1)} \right\} (Y_i^{11})^2 \\
& \quad + 2 \left\{ \frac{m_i^2}{n_i^2} + \chi_{1,1-\alpha}^2 \cdot \frac{m_i (n_i - m_i)}{n_i^2 (n_i - 1)} \right\} Y_i^{00} Y_i^{01} \\
& \quad + 2 \left\{ - \frac{m_i}{n_i} \left(1 - \frac{m_i}{n_i} \right) + \chi_{1,1-\alpha}^2 \cdot \frac{m_i (n_i - m_i)}{n_i^2 (n_i - 1)} \right\} Y_i^{00} Y_i^{10} \\
& \quad + 2 \left\{ - \frac{m_i}{n_i} \left(1 - \frac{m_i}{n_i} \right) + \chi_{1,1-\alpha}^2 \cdot \frac{m_i (n_i - m_i)}{n_i^2 (n_i - 1)} \right\} Y_i^{00} Y_i^{11} \\
& \quad + 2 \left\{ - \frac{m_i}{n_i} \left(1 - \frac{m_i}{n_i} \right) + \chi_{1,1-\alpha}^2 \cdot \frac{m_i (n_i - m_i)}{n_i^2 (n_i - 1)} \right\} Y_i^{01} Y_i^{10}
\end{aligned}$$

$$\begin{aligned}
& + 2 \left\{ -\frac{m_i}{n_i} \left(1 - \frac{m_i}{n_i}\right) + \chi_{1,1-\alpha}^2 \cdot \frac{m_i(n_i - m_i)}{n_i^2(n_i - 1)} \right\} Y_i^{01} Y_i^{11} \\
& + 2 \left\{ \left(1 - \frac{m_i}{n_i}\right)^2 + \chi_{1,1-\alpha}^2 \cdot \frac{m_i(n_i - m_i)}{n_i^2(n_i - 1)} \right\} Y_i^{10} Y_i^{11} \\
& - \chi_{1,1-\alpha}^2 \cdot \frac{m_i n_i (n_i - m_i)}{n_i^2(n_i - 1)} \left(Y_i^{00} + Y_i^{01} + Y_i^{10} + Y_i^{11} \right) \\
& + \sum_{i \neq i'} \left\{ \frac{m_i}{n_i} \frac{m_{i'}}{n_{i'}} Y_i^{00} Y_{i'}^{00} + \frac{m_i}{n_i} \frac{m_{i'}}{n_{i'}} Y_i^{00} Y_{i'}^{01} - \frac{m_i}{n_i} \left(1 - \frac{m_{i'}}{n_{i'}}\right) Y_i^{00} Y_{i'}^{10} \right. \\
& \quad - \frac{m_i}{n_i} \left(1 - \frac{m_{i'}}{n_{i'}}\right) Y_i^{00} Y_{i'}^{11} + \frac{m_i}{n_i} \frac{m_{i'}}{n_{i'}} Y_i^{01} Y_{i'}^{00} + \frac{m_i}{n_i} \frac{m_{i'}}{n_{i'}} Y_i^{01} Y_{i'}^{01} \\
& \quad - \frac{m_i}{n_i} \left(1 - \frac{m_{i'}}{n_{i'}}\right) Y_i^{01} Y_{i'}^{10} - \frac{m_i}{n_i} \left(1 - \frac{m_{i'}}{n_{i'}}\right) Y_i^{01} Y_{i'}^{11} - \left(1 - \frac{m_i}{n_i}\right) \frac{m_{i'}}{n_{i'}} Y_i^{10} Y_{i'}^{00} \\
& \quad - \left(1 - \frac{m_i}{n_i}\right) \frac{m_{i'}}{n_{i'}} Y_i^{10} Y_{i'}^{01} + \left(1 - \frac{m_i}{n_i}\right) \left(1 - \frac{m_{i'}}{n_{i'}}\right) Y_i^{10} Y_{i'}^{10} \\
& \quad + \left(1 - \frac{m_i}{n_i}\right) \left(1 - \frac{m_{i'}}{n_{i'}}\right) Y_i^{10} Y_{i'}^{11} - \left(1 - \frac{m_i}{n_i}\right) \frac{m_{i'}}{n_{i'}} Y_i^{11} Y_{i'}^{00} - \left(1 - \frac{m_i}{n_i}\right) \frac{m_{i'}}{n_{i'}} Y_i^{11} Y_{i'}^{01} \\
& \quad \left. + \left(1 - \frac{m_i}{n_i}\right) \left(1 - \frac{m_{i'}}{n_{i'}}\right) Y_i^{11} Y_{i'}^{10} + \left(1 - \frac{m_i}{n_i}\right) \left(1 - \frac{m_{i'}}{n_{i'}}\right) Y_i^{11} Y_{i'}^{11} \right\},
\end{aligned}$$

which enters into the quadratic constraint for the integer program (P1), of which a standard form is given in Appendix B.1. The above equation can also be written as

$$\begin{aligned}
& [T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y}) - E\{T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y})\}]^2 - \chi_{1,1-\alpha}^2 \cdot \text{Var}\{T_{\text{M-H}}(\mathbf{Z}, \mathbf{Y})\} \\
& = \left\{ \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} (\Delta_{sp}^{10} + \Delta_{sp}^{11}) - \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} \left(\frac{\tilde{m}_s}{\tilde{n}_s} \cdot \check{\Delta}_{sp} \right) \right\}^2 \\
& \quad - \chi_{1,1-\alpha}^2 \cdot \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} \frac{\tilde{m}_s \check{\Delta}_{sp} (\tilde{n}_s - \check{\Delta}_{sp}) (\tilde{n}_s - \tilde{m}_s)}{\tilde{n}_s^2 (\tilde{n}_s - 1)},
\end{aligned}$$

which enters into the quadratic constraint for the integer program (P2), of which a standard form is given in Appendix B.2.

A.3: Some preliminary calculations concerning computing warning accuracy with Neyman's weak null

By Definition 1, if Neyman's weak null hypothesis H_0^{weak} was rejected based on measured outcomes (i.e., $\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y}^*)\}^2 - \chi_{1,1-\alpha}^2 \cdot \widehat{\text{Var}}\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y}^*)\} > 0$), the warning accuracy \mathcal{WA} for testing H_0^{weak} with the Neyman estimator is the optimal value of the following integer quadratically constrained linear program:

$$\begin{aligned} \underset{\mathbf{Y} \in \{0,1\}^N}{\text{maximize}} \quad & \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} Y_{ij}^* Y_{ij} + \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*)(1 - Y_{ij}) \quad (\text{P0}') \\ \text{subject to} \quad & \{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y})\}^2 - \chi_{1,1-\alpha}^2 \cdot \widehat{\text{Var}}\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y})\} \leq 0. \end{aligned}$$

If H_0^{weak} fails to be rejected based on measured outcomes (i.e., $\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y}^*)\}^2 - \chi_{1,1-\alpha}^2 \cdot \widehat{\text{Var}}\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y}^*)\} \leq 0$), we just need to replace the " ≤ 0 " with the " ≥ 0 " in the quadratic constraint in (P0'). As mentioned in the main text, in this paper we will focus on (P0'). For the quadratic constraint in the integer program (P0'), we have

$$\begin{aligned} & \{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y})\}^2 - \chi_{1,1-\alpha}^2 \cdot \widehat{\text{Var}}\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y})\} \\ &= \sum_{i=1}^I \sum_{j=1}^{n_i} \left[\left\{ \frac{n_i}{Nm_i} Z_{ij} - \frac{n_i}{N(n_i - m_i)} (1 - Z_{ij}) \right\}^2 \right. \\ & \quad - \chi_{1,1-\alpha}^2 \cdot \frac{\left(\frac{n_i}{N}\right)^2}{m_i(m_i - 1)} Z_{ij} \left(1 - \frac{Z_{ij}}{m_i}\right)^2 \\ & \quad - \chi_{1,1-\alpha}^2 \cdot \frac{\left(\frac{n_i}{N}\right)^2}{m_i(m_i - 1)} \sum_{j' \neq j} \frac{Z_{ij'} Z_{ij}^2}{m_i^2} \\ & \quad - \chi_{1,1-\alpha}^2 \cdot \frac{\left(\frac{n_i}{N}\right)^2}{(n_i - m_i)(n_i - m_i - 1)} (1 - Z_{ij}) \left(1 - \frac{1 - Z_{ij}}{n_i - m_i}\right)^2 \\ & \quad \left. - \chi_{1,1-\alpha}^2 \cdot \frac{\left(\frac{n_i}{N}\right)^2}{(n_i - m_i)(n_i - m_i - 1)} \sum_{j' \neq j} \frac{(1 - Z_{ij'})(1 - Z_{ij})^2}{(n_i - m_i)^2} \right] Y_{ij}^2 \end{aligned}$$

$$\begin{aligned}
& + \sum_{i=1}^I \sum_{j \neq j'} \left[\left\{ \frac{n_i}{Nm_i} Z_{ij} - \frac{n_i}{N(n_i - m_i)} (1 - Z_{ij}) \right\} \left\{ \frac{n_i}{Nm_i} Z_{ij'} - \frac{n_i}{N(n_i - m_i)} (1 - Z_{ij'}) \right\} \right. \\
& \quad + \chi_{1,1-\alpha}^2 \cdot \frac{\left(\frac{n_i}{N}\right)^2}{m_i(m_i - 1)} Z_{ij} \left(1 - \frac{Z_{ij}}{m_i}\right) \frac{Z_{ij'}}{m_i} \\
& \quad + \chi_{1,1-\alpha}^2 \cdot \frac{\left(\frac{n_i}{N}\right)^2}{m_i(m_i - 1)} Z_{ij'} \left(1 - \frac{Z_{ij'}}{m_i}\right) \frac{Z_{ij}}{m_i} \\
& \quad - \chi_{1,1-\alpha}^2 \cdot \frac{\left(\frac{n_i}{N}\right)^2}{m_i(m_i - 1)} \sum_{j'' \neq j, j'} \frac{Z_{ij''} Z_{ij} Z_{ij'}}{m_i^2} \\
& \quad + \chi_{1,1-\alpha}^2 \cdot \frac{\left(\frac{n_i}{N}\right)^2}{(n_i - m_i)(n_i - m_i - 1)} (1 - Z_{ij}) \left(1 - \frac{1 - Z_{ij}}{n_i - m_i}\right) \frac{1 - Z_{ij'}}{n_i - m_i} \\
& \quad + \chi_{1,1-\alpha}^2 \cdot \frac{\left(\frac{n_i}{N}\right)^2}{(n_i - m_i)(n_i - m_i - 1)} (1 - Z_{ij'}) \left(1 - \frac{1 - Z_{ij'}}{n_i - m_i}\right) \frac{1 - Z_{ij}}{n_i - m_i} \\
& \quad \left. - \chi_{1,1-\alpha}^2 \cdot \frac{\left(\frac{n_i}{N}\right)^2}{(n_i - m_i)(n_i - m_i - 1)} \sum_{j'' \neq j, j'} \frac{(1 - Z_{ij''})(1 - Z_{ij})(1 - Z_{ij'})}{(n_i - m_i)^2} \right] Y_{ij} Y_{ij'} \\
& + \sum_{i \neq i'} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_{i'}} \left\{ \frac{n_i}{Nm_i} Z_{ij} - \frac{n_i}{N(n_i - m_i)} (1 - Z_{ij}) \right\} \\
& \quad \times \left\{ \frac{n_{i'}}{Nm_{i'}} Z_{i'j'} - \frac{n_{i'}}{N(n_{i'} - m_{i'})} (1 - Z_{i'j'}) \right\} Y_{ij} Y_{i'j'}.
\end{aligned}$$

The above equation can be rewritten as

$$\begin{aligned}
& \{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y})\}^2 - \chi_{1,1-\alpha}^2 \cdot \widehat{\text{Var}}\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y})\} \\
& = \left[\sum_{i=1}^I \frac{n_i}{N} \left\{ \frac{1}{m_i} \sum_{j=1}^{n_i} Z_{ij} Y_{ij} - \frac{1}{n_i - m_i} \sum_{j=1}^{n_i} (1 - Z_{ij}) Y_{ij} \right\} \right]^2 \\
& \quad - \chi_{1,1-\alpha}^2 \cdot \sum_{i=1}^I \left(\frac{n_i}{N}\right)^2 \left[\frac{1}{m_i(m_i - 1)} \sum_{j=1}^{n_i} Z_{ij} \left(Y_{ij} - \frac{1}{m_i} \sum_{j'=1}^{n_i} Z_{ij'} Y_{ij'}\right)^2 \right. \\
& \quad \left. + \frac{1}{(n_i - m_i)(n_i - m_i - 1)} \sum_{j=1}^{n_i} (1 - Z_{ij}) \left\{ Y_{ij} - \frac{1}{n_i - m_i} \sum_{j'=1}^{n_i} (1 - Z_{ij'}) Y_{ij'} \right\}^2 \right] \\
& = \left[\sum_{i=1}^I \left\{ \frac{n_i}{Nm_i} (Y_i^{10} + Y_i^{11}) - \frac{n_i}{N(n_i - m_i)} (Y_i^{00} + Y_i^{01}) \right\} \right]^2 \\
& \quad - \chi_{1,1-\alpha}^2 \cdot \sum_{i=1}^I \left(\frac{n_i}{N}\right)^2 \left\{ \frac{Y_i^{10} + Y_i^{11}}{m_i(m_i - 1)} - \frac{(Y_i^{10} + Y_i^{11})^2}{m_i^2(m_i - 1)} \right\}
\end{aligned}$$

$$\begin{aligned}
& + \frac{Y_i^{00} + Y_i^{01}}{(n_i - m_i)(n_i - m_i - 1)} - \frac{(Y_i^{00} + Y_i^{01})^2}{(n_i - m_i)^2(n_i - m_i - 1)} \Big\} \\
= & \sum_{i=1}^I \left[\left\{ \frac{n_i^2}{N^2(n_i - m_i)^2} + \chi_{1,1-\alpha}^2 \cdot \frac{n_i^2}{N^2(n_i - m_i)^2(n_i - m_i - 1)} \right\} \left\{ (Y_i^{00})^2 + (Y_i^{01})^2 \right\} \right. \\
& + \left\{ \frac{n_i^2}{N^2 m_i^2} + \chi_{1,1-\alpha}^2 \cdot \frac{n_i^2}{N^2 m_i^2(m_i - 1)} \right\} \left\{ (Y_i^{10})^2 + (Y_i^{11})^2 \right\} \\
& + 2 \left\{ \frac{n_i^2}{N^2(n_i - m_i)^2} + \chi_{1,1-\alpha}^2 \cdot \frac{n_i^2}{N^2(n_i - m_i)^2(n_i - m_i - 1)} \right\} Y_i^{00} Y_i^{01} \\
& - \frac{2n_i^2}{N^2 m_i(n_i - m_i)} Y_i^{00} Y_i^{10} - \frac{2n_i^2}{N^2 m_i(n_i - m_i)} Y_i^{00} Y_i^{11} \\
& - \frac{2n_i^2}{N^2 m_i(n_i - m_i)} Y_i^{01} Y_i^{10} - \frac{2n_i^2}{N^2 m_i(n_i - m_i)} Y_i^{01} Y_i^{11} \\
& + 2 \left\{ \frac{n_i^2}{N^2 m_i^2} + \chi_{1,1-\alpha}^2 \cdot \frac{n_i^2}{N^2 m_i^2(m_i - 1)} \right\} Y_i^{10} Y_i^{11} \\
& - \frac{\chi_{1,1-\alpha}^2 \cdot n_i^2}{N^2(n_i - m_i)(n_i - m_i - 1)} (Y_i^{00} + Y_i^{01}) - \frac{\chi_{1,1-\alpha}^2 \cdot n_i^2}{N^2 m_i(m_i - 1)} (Y_i^{10} + Y_i^{11}) \Big] \\
& + \sum_{i \neq i'} \left\{ \frac{n_i n_{i'}}{N^2(n_i - m_i)(n_{i'} - m_{i'})} Y_i^{00} Y_{i'}^{00} + \frac{n_i n_{i'}}{N^2(n_i - m_i)(n_{i'} - m_{i'})} Y_i^{00} Y_{i'}^{01} \right. \\
& - \frac{n_i n_{i'}}{N^2(n_i - m_i) m_{i'}} Y_i^{00} Y_{i'}^{10} - \frac{n_i n_{i'}}{N^2(n_i - m_i) m_{i'}} Y_i^{00} Y_{i'}^{11} \\
& + \frac{n_i n_{i'}}{N^2(n_i - m_i)(n_{i'} - m_{i'})} Y_i^{01} Y_{i'}^{00} + \frac{n_i n_{i'}}{N^2(n_i - m_i)(n_{i'} - m_{i'})} Y_i^{01} Y_{i'}^{01} \\
& - \frac{n_i n_{i'}}{N^2(n_i - m_i) m_{i'}} Y_i^{01} Y_{i'}^{10} - \frac{n_i n_{i'}}{N^2(n_i - m_i) m_{i'}} Y_i^{01} Y_{i'}^{11} \\
& - \frac{n_i n_{i'}}{N^2 m_i(n_{i'} - m_{i'})} Y_i^{10} Y_{i'}^{00} - \frac{n_i n_{i'}}{N^2 m_i(n_{i'} - m_{i'})} Y_i^{10} Y_{i'}^{01} \\
& + \frac{n_i n_{i'}}{N^2 m_i m_{i'}} Y_i^{10} Y_{i'}^{10} + \frac{n_i n_{i'}}{N^2 m_i m_{i'}} Y_i^{10} Y_{i'}^{11} - \frac{n_i n_{i'}}{N^2 m_i(n_{i'} - m_{i'})} Y_i^{11} Y_{i'}^{00} \\
& \left. - \frac{n_i n_{i'}}{N^2 m_i(n_{i'} - m_{i'})} Y_i^{11} Y_{i'}^{01} + \frac{n_i n_{i'}}{N^2 m_i m_{i'}} Y_i^{11} Y_{i'}^{10} + \frac{n_i n_{i'}}{N^2 m_i m_{i'}} Y_i^{11} Y_{i'}^{11} \right\},
\end{aligned}$$

which enters into the quadratic constraint for the integer program (P3), of which a standard form is given in Appendix B.3. The above equation can also be written as

$$\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y})\}^2 - \chi_{1,1-\alpha}^2 \cdot \widehat{\text{Var}}\{T_{\text{Neyman}}(\mathbf{Z}, \mathbf{Y})\}$$

$$\begin{aligned}
&= \left[\sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} \left\{ \frac{\tilde{n}_s}{N\tilde{m}_s} (\Delta_{sp}^{10} + \Delta_{sp}^{11}) - \frac{\tilde{n}_s}{N(\tilde{n}_s - \tilde{m}_s)} (\Delta_{sp}^{00} + \Delta_{sp}^{01}) \right\} \right]^2 \\
&\quad - \chi_{1,1-\alpha}^2 \cdot \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} \left(\frac{\tilde{n}_s}{N} \right)^2 \left\{ \frac{\Delta_{sp}^{10} + \Delta_{sp}^{11}}{\tilde{m}_s(\tilde{m}_s - 1)} - \frac{(\Delta_{sp}^{10} + \Delta_{sp}^{11})^2}{\tilde{m}_s^2(\tilde{m}_s - 1)} \right. \\
&\quad \left. + \frac{\Delta_{sp}^{00} + \Delta_{sp}^{01}}{(\tilde{n}_s - \tilde{m}_s)(\tilde{n}_s - \tilde{m}_s - 1)} - \frac{(\Delta_{sp}^{00} + \Delta_{sp}^{01})^2}{(\tilde{n}_s - \tilde{m}_s)^2(\tilde{n}_s - \tilde{m}_s - 1)} \right\},
\end{aligned}$$

which enters into the quadratic constraint for the integer program (P4), of which a standard form is given in Appendix B.4.

Appendix B: Detailed Formulations of the Related Integer Programs for Computing Warning Accuracy in Various Cases

B.1: Warning accuracy with Fisher's sharp null (type I randomization designs)

We write the following integer quadratically constrained linear program

$$\begin{aligned}
& \max_{\mathbf{Y} \in \mathbb{Z}^{4I}} \quad \frac{1}{N} \sum_{i=1}^I (Y_i^{01} + Y_i^{11} - Y_i^{00} - Y_i^{10}) + \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*) \quad (\text{P1}) \\
& \text{s.t.} \quad \left\{ \sum_{i=1}^I (Y_i^{10} + Y_i^{11}) - \sum_{i=1}^I \frac{m_i}{n_i} \check{Y}_i \right\}^2 - \chi_{1,1-\alpha}^2 \cdot \sum_{i=1}^I \frac{m_i \check{Y}_i (n_i - \check{Y}_i) (n_i - m_i)}{n_i^2 (n_i - 1)} \leq 0, \\
& \quad 0 \leq Y_i^{00} \leq \sum_{j=1}^{n_i} (1 - Z_{ij}) (1 - Y_{ij}^*), \quad \forall i \\
& \quad 0 \leq Y_i^{01} \leq \sum_{j=1}^{n_i} (1 - Z_{ij}) Y_{ij}^*, \quad \forall i \\
& \quad 0 \leq Y_i^{10} \leq \sum_{j=1}^{n_i} Z_{ij} (1 - Y_{ij}^*), \quad \forall i \\
& \quad 0 \leq Y_i^{11} \leq \sum_{j=1}^{n_i} Z_{ij} Y_{ij}^*, \quad \forall i
\end{aligned}$$

in a standard form

$$\begin{aligned}
& \max_{\mathbf{x}} \quad \mathbf{q}^T \mathbf{x} + c \\
& \text{s.t.} \quad \mathbf{x}^T \mathbf{Q}_1 \mathbf{x} + \mathbf{q}_1^T \mathbf{x} \leq 0, \\
& \quad \mathbf{1} \leq \mathbf{x} \leq \mathbf{u},
\end{aligned}$$

All elements of \mathbf{x} are integers.

Specially, we have

- decision variables: $\mathbf{x} = \mathbf{Y} = (Y_1^{00}, Y_1^{01}, Y_1^{10}, Y_1^{11}, \dots, Y_I^{00}, Y_I^{01}, Y_I^{10}, Y_I^{11})$.
- objective function: $\mathbf{q}^T \mathbf{x} + c$ where

$$\mathbf{q} = \left(-\frac{1}{N'}, \frac{1}{N'}, -\frac{1}{N'}, \frac{1}{N'}, \dots, -\frac{1}{N'}, \frac{1}{N'}, -\frac{1}{N'}, \frac{1}{N'} \right) \text{ and } c = \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*).$$

- quadratic constraint: $\mathbf{x}^T \mathbf{Q}_1 \mathbf{x} + \mathbf{q}_1^T \mathbf{x} \leq 0$ where $\mathbf{Q}_1 = (Q_{1,s,t})_{4I \times 4I}$ is a $4I \times 4I$ matrix. Suppose that $s = 4(i-1) + k$ and $t = 4(i'-1) + k'$ for some integers $i, i' \in \{1, \dots, I\}$ and $k, k' \in \{1, 2, 3, 4\}$. Then we have:

1. If (s, t) satisfies one of the following conditions: 1) $i = i'$ and $k = k' = 1$; 2) $i = i'$ and $k = k' = 2$; 3) $i = i'$ and $k = 1, k' = 2$; 4) $i = i'$ and $k = 2, k' = 1$, we have

$$Q_{1,s,t} = \frac{m_i^2}{n_i^2} + \chi_{1,1-\alpha}^2 \cdot \frac{m_i(n_i - m_i)}{n_i^2(n_i - 1)}.$$

2. If (s, t) satisfies one of the following conditions: 1) $i = i'$ and $k = k' = 3$; 2) $i = i'$ and $k = k' = 4$; 3) $i = i'$ and $k = 3, k' = 4$; 4) $i = i'$ and $k = 4, k' = 3$, we have

$$Q_{1,s,t} = \left(1 - \frac{m_i}{n_i}\right)^2 + \chi_{1,1-\alpha}^2 \cdot \frac{m_i(n_i - m_i)}{n_i^2(n_i - 1)}.$$

3. If (s, t) satisfies one of the following conditions: 1) $i = i'$ and $k = 1, k' = 3$; 2) $i = i'$ and $k = 3, k' = 1$; 3) $i = i'$ and $k = 1, k' = 4$; 4) $i = i'$ and $k = 4, k' = 1$; 5) $i = i'$ and $k = 2, k' = 3$; 6) $i = i'$ and $k = 3, k' = 2$; 7)

$i = i'$ and $k = 2, k' = 4$; 8) $i = i'$ and $k = 4, k' = 2$; we have

$$Q_{1,s,t} = -\frac{m_i}{n_i} \left(1 - \frac{m_i}{n_i}\right) + \chi_{1,1-\alpha}^2 \cdot \frac{m_i(n_i - m_i)}{n_i^2(n_i - 1)}.$$

4. If (s, t) satisfies one of the following conditions: 1) $i \neq i'$ and $k = k' = 1$; 2) $i \neq i'$ and $k = 1, k' = 2$; 3) $i \neq i'$ and $k = 2, k' = 1$; 4) $i \neq i'$ and $k = k' = 2$, we have

$$Q_{1,s,t} = \frac{m_i}{n_i} \frac{m_{i'}}{n_{i'}}.$$

5. If (s, t) satisfies one of the following conditions: 1) $i \neq i'$ and $k = 1, k' = 3$; 2) $i \neq i'$ and $k = 1, k' = 4$; 3) $i \neq i'$ and $k = 2, k' = 3$; 4) $i \neq i'$ and $k = 2, k' = 4$, we have

$$Q_{1,s,t} = -\frac{m_i}{n_i} \left(1 - \frac{m_{i'}}{n_{i'}}\right).$$

6. If (s, t) satisfies one of the following conditions: 1) $i \neq i'$ and $k = 3, k' = 1$; 2) $i \neq i'$ and $k = 3, k' = 2$; 3) $i \neq i'$ and $k = 4, k' = 1$; 4) $i \neq i'$ and $k = 4, k' = 2$, we have

$$Q_{1,s,t} = -\left(1 - \frac{m_i}{n_i}\right) \frac{m_{i'}}{n_{i'}}.$$

7. If (s, t) satisfies one of the following conditions: 1) $i \neq i'$ and $k = 3, k' = 3$; 2) $i \neq i'$ and $k = 3, k' = 4$; 3) $i \neq i'$ and $k = 4, k' = 3$; 4) $i \neq i'$ and $k = 4, k' = 4$, we have

$$Q_{1,s,t} = \left(1 - \frac{m_i}{n_i}\right) \left(1 - \frac{m_{i'}}{n_{i'}}\right).$$

We have $\mathbf{q}_1 = (q_{1,1}, \dots, q_{1,4I})$ is a $4I$ -dimensional vector where

$$q_{1,s} = -\chi_{1,1-\alpha}^2 \cdot \frac{m_i n_i (n_i - m_i)}{n_i^2 (n_i - 1)}, \quad \text{for } s = 1, \dots, 4I.$$

- box constraints: $\mathbf{l} \leq \mathbf{Y} \leq \mathbf{u}$, where $\mathbf{l} = \mathbf{0}$ and $\mathbf{M} = (M_1^{00}, M_1^{01}, M_1^{10}, M_1^{11}, \dots, M_I^{00}, M_I^{01}, M_I^{10}, M_I^{11})$ is a $4I$ -dimensional vector with $M_i^{00} = \sum_{j=1}^{n_i} (1 - Z_{ij})(1 - Y_{ij}^*)$, $M_i^{01} = \sum_{j=1}^{n_i} (1 - Z_{ij})Y_{ij}^*$, $M_i^{10} = \sum_{j=1}^{n_i} Z_{ij}(1 - Y_{ij}^*)$, and $M_i^{11} = \sum_{j=1}^{n_i} Z_{ij}Y_{ij}^*$.
- integrality constraints: all $4I$ elements of \mathbf{x} are integers.

B.2: Warning accuracy with Fisher's sharp null (type II randomization designs)

We write the following integer quadratically constrained linear program

$$\begin{aligned} \max_{d_{sp} \in \mathbb{Z}} \quad & \frac{1}{N} \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} (\Delta_{sp}^{01} + \Delta_{sp}^{11} - \Delta_{sp}^{00} - \Delta_{sp}^{10}) + \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*) \quad (\text{P2}) \\ \text{s.t.} \quad & \left\{ \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} (\Delta_{sp}^{10} + \Delta_{sp}^{11}) - \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} \left(\frac{\tilde{m}_s}{\tilde{n}_s} \cdot \check{\Delta}_{sp} \right) \right\}^2 \\ & - \chi_{1,1-\alpha}^2 \cdot \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} \frac{\tilde{m}_s \check{\Delta}_{sp} (\tilde{n}_s - \check{\Delta}_{sp}) (\tilde{n}_s - \tilde{m}_s)}{\tilde{n}_s^2 (\tilde{n}_s - 1)} \leq 0, \\ & \sum_{p=1}^{\tilde{N}_s} d_{sp} = P_s, \quad \forall s \\ & d_{sp} \geq 0, \quad \forall s, p \end{aligned}$$

in a standard form

$$\begin{aligned}
& \max_{\mathbf{x}} \quad \mathbf{q}^T \mathbf{x} + c \\
& \text{s.t.} \quad \mathbf{x}^T \mathbf{Q}_1 \mathbf{x} + \mathbf{q}_1^T \mathbf{x} \leq 0, \\
& \quad \quad \mathbf{q}_{2s}^T \mathbf{x} = P_s, \quad \forall s \\
& \quad \quad \mathbf{x} \geq \mathbf{0}, \\
& \quad \quad \text{All elements of } \mathbf{x} \text{ are integers.}
\end{aligned}$$

We have:

- decision variables: $\mathbf{x} = \mathbf{d} = (d_{11}, \dots, d_{S\tilde{N}_S})$;
- objective function: $\mathbf{q}^T \mathbf{x} + c$ where

$$\mathbf{q} = \left(\frac{\Delta_{11}^{01} + \Delta_{11}^{11} - \Delta_{11}^{00} - \Delta_{11}^{10}}{N}, \dots, \frac{\Delta_{S\tilde{N}_S}^{01} + \Delta_{S\tilde{N}_S}^{11} - \Delta_{S\tilde{N}_S}^{00} - \Delta_{S\tilde{N}_S}^{10}}{N} \right),$$

and $c = \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*)$.

- quadratic constraint: $\mathbf{x}^T \mathbf{Q}_1 \mathbf{x} + \mathbf{q}_1^T \mathbf{x} \leq 0$ where $\mathbf{Q}_1 = (Q_{1,r,t})_{\tilde{N} \times \tilde{N}}$ is a $\tilde{N} \times \tilde{N}$ matrix ($\tilde{N} = \sum_{s=1}^S \tilde{N}_s$). Suppose that r corresponds to the p -th unique 2×2 table for the s -th unique 2×2 table $\Lambda_{[s]}$, and t corresponds to the p' -th unique $2 \times 2 \times 2$ table for the s' -th unique 2×2 table $\Lambda_{[s']}$. Then we have:

$$\begin{aligned}
Q_{1,r,t} = & \left\{ \Delta_{sp}^{10} + \Delta_{sp}^{11} - \frac{\tilde{m}_s}{\tilde{n}_s} (\Delta_{sp}^{00} + \Delta_{sp}^{01} + \Delta_{sp}^{10} + \Delta_{sp}^{11}) \right\} \\
& \times \left\{ \Delta_{s'p'}^{10} + \Delta_{s'p'}^{11} - \frac{\tilde{m}_{s'}}{\tilde{n}_{s'}} (\Delta_{s'p'}^{00} + \Delta_{s'p'}^{01} + \Delta_{s'p'}^{10} + \Delta_{s'p'}^{11}) \right\}.
\end{aligned}$$

We have $\mathbf{q}_1 = (q_{11}, \dots, q_{S\tilde{N}_s})$ is a \tilde{N} -dimensional vector where

$$q_{1,s,p} = -\chi_{1,1-\alpha}^2 \cdot \frac{\tilde{m}_s \check{\Delta}_{sp} (\tilde{n}_s - \check{\Delta}_{sp}) (\tilde{n}_s - \tilde{m}_s)}{\tilde{n}_s^2 (\tilde{n}_s - 1)}.$$

- linear constraints: $\mathbf{q}_{2s}^T \mathbf{x} = P_s$, where \mathbf{q}_{2s} is the zero-one indicator vector for all the \tilde{N}_s possible unique $2 \times 2 \times 2$ tables of $\Lambda_{[s]}$.
- box constraints: $d_{sp} \geq 0$ for all s and p .
- integrality constraints: all \tilde{N} elements of \mathbf{x} are integers.

B.3: Warning accuracy with Neyman's weak null (type I randomization designs)

Following a similar argument as in Section 3.4.2.1, we can reformulate the integer program (P0') as the following integer quadratically constrained linear program

$$\begin{aligned}
\max_{\mathbf{Y} \in \mathbb{Z}^{4I}} \quad & \frac{1}{N} \sum_{i=1}^I (Y_i^{01} + Y_i^{11} - Y_i^{00} - Y_i^{10}) + \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*) \quad (\text{P3}) \\
\text{s.t.} \quad & \left[\sum_{i=1}^I \left\{ \frac{n_i}{Nm_i} (Y_i^{10} + Y_i^{11}) - \frac{n_i}{N(n_i - m_i)} (Y_i^{00} + Y_i^{01}) \right\} \right]^2 \\
& - \chi_{1,1-\alpha}^2 \cdot \sum_{i=1}^I \left(\frac{n_i}{N} \right)^2 \left\{ \frac{Y_i^{10} + Y_i^{11}}{m_i(m_i - 1)} - \frac{(Y_i^{10} + Y_i^{11})^2}{m_i^2(m_i - 1)} \right. \\
& \quad \left. + \frac{Y_i^{00} + Y_i^{01}}{(n_i - m_i)(n_i - m_i - 1)} - \frac{(Y_i^{00} + Y_i^{01})^2}{(n_i - m_i)^2(n_i - m_i - 1)} \right\} \leq 0, \\
& 0 \leq Y_i^{00} \leq \sum_{j=1}^{n_i} (1 - Z_{ij})(1 - Y_{ij}^*), \quad \forall i \\
& 0 \leq Y_i^{01} \leq \sum_{j=1}^{n_i} (1 - Z_{ij})Y_{ij}^*, \quad \forall i \\
& 0 \leq Y_i^{10} \leq \sum_{j=1}^{n_i} Z_{ij}(1 - Y_{ij}^*), \quad \forall i \\
& 0 \leq Y_i^{11} \leq \sum_{j=1}^{n_i} Z_{ij}Y_{ij}^*, \quad \forall i
\end{aligned}$$

which can be written as the following standard form

$$\begin{aligned}
\max_{\mathbf{x}} \quad & \mathbf{q}^T \mathbf{x} + c \\
\text{s.t.} \quad & \mathbf{x}^T \mathbf{Q}_1 \mathbf{x} + \mathbf{q}_1^T \mathbf{x} \leq 0, \\
& \mathbf{l} \leq \mathbf{x} \leq \mathbf{u},
\end{aligned}$$

All elements of \mathbf{x} are integers.

Specifically, we have:

- decision variables: $\mathbf{x} = \mathbf{Y} = (Y_1^{00}, Y_1^{01}, Y_1^{10}, Y_1^{11}, \dots, Y_I^{00}, Y_I^{01}, Y_I^{10}, Y_I^{11})$;
- objective function: $\mathbf{q}^T \mathbf{x} + c$ where

$$\mathbf{q} = \left(-\frac{1}{N'} \frac{1}{N'}, -\frac{1}{N'} \frac{1}{N'}, \dots, -\frac{1}{N'} \frac{1}{N'}, -\frac{1}{N'} \frac{1}{N'} \right) \text{ and } c = \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*).$$

- quadratic constraint: $\mathbf{x}^T \mathbf{Q}_1 \mathbf{x} + \mathbf{q}_1^T \mathbf{x} \leq 0$ where $\mathbf{Q}_1 = (Q_{1,s,t})_{4I \times 4I}$ is a $4I \times 4I$ matrix. Suppose that $s = 4(i-1) + k$ and $t = 4(i'-1) + k'$ for some integers $i, i' \in \{1, \dots, I\}$ and $k, k' \in \{1, 2, 3, 4\}$. Then we have:

1. If (s, t) satisfies one of the following conditions: 1) $i = i'$ and $k = k' = 1$; 2) $i = i'$ and $k = k' = 2$; 3) $i = i'$ and $k = 1, k' = 2$; 4) $i = i'$ and $k = 2, k' = 1$, we have

$$Q_{1,s,t} = \frac{n_i^2}{N^2(n_i - m_i)^2} + \chi_{1,1-\alpha}^2 \cdot \frac{n_i^2}{N^2(n_i - m_i)^2(n_i - m_i - 1)}.$$

2. If (s, t) satisfies one of the following conditions: 1) $i = i'$ and $k = k' = 3$; 2) $i = i'$ and $k = k' = 4$; 3) $i = i'$ and $k = 3, k' = 4$; 4) $i = i'$ and $k = 4, k' = 3$, we have

$$Q_{1,s,t} = \frac{n_i^2}{N^2 m_i^2} + \chi_{1,1-\alpha}^2 \cdot \frac{n_i^2}{N^2 m_i^2 (m_i - 1)}.$$

3. If (s, t) satisfies one of the following conditions: 1) $i = i'$ and $k = 1, k' = 3$; 2) $i = i'$ and $k = 3, k' = 1$; 3) $i = i'$ and $k = 1, k' = 4$; 4) $i = i'$ and $k = 4, k' = 1$; 5) $i = i'$ and $k = 2, k' = 3$; 6) $i = i'$ and $k = 3, k' = 2$; 7)

$i = i'$ and $k = 2, k' = 4$; 8) $i = i'$ and $k = 4, k' = 2$; we have

$$Q_{1,s,t} = -\frac{n_i^2}{N^2 m_i (n_i - m_i)}.$$

4. If (s, t) satisfies one of the following conditions: 1) $i \neq i'$ and $k = k' = 1$; 2) $i \neq i'$ and $k = 1, k' = 2$; 3) $i \neq i'$ and $k = 2, k' = 1$; 4) $i \neq i'$ and $k = k' = 2$, we have

$$Q_{1,s,t} = \frac{n_i n_{i'}}{N^2 (n_i - m_i) (n_{i'} - m_{i'})}.$$

5. If (s, t) satisfies one of the following conditions: 1) $i \neq i'$ and $k = 1, k' = 3$; 2) $i \neq i'$ and $k = 1, k' = 4$; 3) $i \neq i'$ and $k = 2, k' = 3$; 4) $i \neq i'$ and $k = 2, k' = 4$, we have

$$Q_{1,s,t} = -\frac{n_i n_{i'}}{N^2 (n_i - m_i) m_{i'}}.$$

6. If (s, t) satisfies one of the following conditions: 1) $i \neq i'$ and $k = 3, k' = 1$; 2) $i \neq i'$ and $k = 3, k' = 2$; 3) $i \neq i'$ and $k = 4, k' = 1$; 4) $i \neq i'$ and $k = 4, k' = 2$, we have

$$Q_{1,s,t} = -\frac{n_i n_{i'}}{N^2 m_i (n_{i'} - m_{i'})}.$$

7. If (s, t) satisfies one of the following conditions: 1) $i \neq i'$ and $k = 3, k' = 3$; 2) $i \neq i'$ and $k = 3, k' = 4$; 3) $i \neq i'$ and $k = 4, k' = 3$; 4) $i \neq i'$ and

$k = 4, k' = 4$, we have

$$Q_{1,s,t} = \frac{n_i n_{i'}}{N^2 m_i m_{i'}}.$$

We have $\mathbf{q}_1 = (q_{1,1}, \dots, q_{1,4I})$ is a $4I$ -dimensional vector where

$$q_{1,s} = -\frac{\chi_{1,1-\alpha}^2 \cdot n_i^2}{N^2 (n_i - m_i)(n_i - m_i - 1)} \text{ for } k = 1, 2,$$

and

$$q_{1,s} = -\frac{\chi_{1,1-\alpha}^2 \cdot n_i^2}{N^2 m_i (m_i - 1)} \text{ for } k = 3, 4.$$

- box constraints: $\mathbf{l} \leq \mathbf{Y} \leq \mathbf{u}$, where $\mathbf{l} = \mathbf{0}$ and $\mathbf{M} = (M_1^{00}, M_1^{01}, M_1^{10}, M_1^{11}, \dots, M_I^{00}, M_I^{01}, M_I^{10}, M_I^{11})$ is a $4I$ -dimensional vector with $M_i^{00} = \sum_{j=1}^{n_i} (1 - Z_{ij})(1 - Y_{ij}^*)$, $M_i^{01} = \sum_{j=1}^{n_i} (1 - Z_{ij})Y_{ij}^*$, $M_i^{10} = \sum_{j=1}^{n_i} Z_{ij}(1 - Y_{ij}^*)$, and $M_i^{11} = \sum_{j=1}^{n_i} Z_{ij}Y_{ij}^*$.
- integrality constraints: all $4I$ elements of \mathbf{x} are integers.

B.4: Warning accuracy with Neyman's weak null (type II randomization designs)

Following a similar argument as in Section 3.4.2.1, we can reformulate the integer program (P0') as the following integer quadratically constrained linear program

$$\begin{aligned}
\max_{d_{sp} \in \mathbb{Z}} \quad & \frac{1}{N} \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} (\Delta_{sp}^{01} + \Delta_{sp}^{11} - \Delta_{sp}^{00} - \Delta_{sp}^{10}) + \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*) \quad (\text{P4}) \\
\text{s.t.} \quad & \left[\sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} \left\{ \frac{\tilde{n}_s}{N \tilde{m}_s} (\Delta_{sp}^{10} + \Delta_{sp}^{11}) - \frac{\tilde{n}_s}{N(\tilde{n}_s - \tilde{m}_s)} (\Delta_{sp}^{00} + \Delta_{sp}^{01}) \right\} \right]^2 \\
& - \chi_{1,1-\alpha}^2 \cdot \sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} d_{sp} \left(\frac{\tilde{n}_s}{N} \right)^2 \left\{ \frac{\Delta_{sp}^{10} + \Delta_{sp}^{11}}{\tilde{m}_s(\tilde{m}_s - 1)} - \frac{(\Delta_{sp}^{10} + \Delta_{sp}^{11})^2}{\tilde{m}_s^2(\tilde{m}_s - 1)} \right. \\
& \quad \left. + \frac{\Delta_{sp}^{00} + \Delta_{sp}^{01}}{(\tilde{n}_s - \tilde{m}_s)(\tilde{n}_s - \tilde{m}_s - 1)} - \frac{(\Delta_{sp}^{00} + \Delta_{sp}^{01})^2}{(\tilde{n}_s - \tilde{m}_s)^2(\tilde{n}_s - \tilde{m}_s - 1)} \right\} \leq 0, \\
& \sum_{p=1}^{\tilde{N}_s} d_{sp} = P_s, \quad \forall s \\
& d_{sp} \geq 0, \quad \forall s, p
\end{aligned}$$

which can be written as the following standard form

$$\begin{aligned}
\max_{\mathbf{x}} \quad & \mathbf{q}^T \mathbf{x} + c \\
\text{s.t.} \quad & \mathbf{x}^T \mathbf{Q}_1 \mathbf{x} + \mathbf{q}_1^T \mathbf{x} \leq 0, \\
& \mathbf{q}_{2s}^T \mathbf{x} = P_s, \quad \forall s \\
& \mathbf{x} \geq \mathbf{0},
\end{aligned}$$

All elements of \mathbf{x} are integers.

Specifically, we have:

- decision variables: $\mathbf{x} = \mathbf{d} = (d_{11}, \dots, d_{S\tilde{N}_S})$;

- objective function: $\mathbf{q}^T \mathbf{x} + c$ where

$$\mathbf{q} = \left(\frac{\Delta_{11}^{01} + \Delta_{11}^{11} - \Delta_{11}^{00} - \Delta_{11}^{10}}{N}, \dots, \frac{\Delta_{S\tilde{N}_s}^{01} + \Delta_{S\tilde{N}_s}^{11} - \Delta_{S\tilde{N}_s}^{00} - \Delta_{S\tilde{N}_s}^{10}}{N} \right),$$

and

$$c = \frac{1}{N} \sum_{i=1}^I \sum_{j=1}^{n_i} (1 - Y_{ij}^*).$$

- quadratic constraint: $\mathbf{x}^T \mathbf{Q}_1 \mathbf{x} + \mathbf{q}_1^T \mathbf{x} \leq 0$ where $\mathbf{Q}_1 = (Q_{1,r,t})_{\tilde{N} \times \tilde{N}}$ is a $\tilde{N} \times \tilde{N}$ matrix ($\tilde{N} = \sum_{s=1}^S \tilde{N}_s$). Suppose that r corresponds to the p -th unique $2 \times 2 \times 2$ table for the s -th unique 2×2 table $\Lambda_{[s]}$, and t corresponds to the p' -th unique $2 \times 2 \times 2$ table for the s' -th unique 2×2 table $\Lambda_{[s']}$. Then we have

$$Q_{1,r,t} = \left\{ \frac{\tilde{n}_s}{N\tilde{m}_s} (\Delta_{sp}^{10} + \Delta_{sp}^{11}) - \frac{\tilde{n}_s}{N(\tilde{n}_s - \tilde{m}_s)} (\Delta_{sp}^{00} + \Delta_{sp}^{01}) \right\} \\ \times \left\{ \frac{\tilde{n}_{s'}}{N\tilde{m}_{s'}} (\Delta_{s'p'}^{10} + \Delta_{s'p'}^{11}) - \frac{\tilde{n}_{s'}}{N(\tilde{n}_{s'} - \tilde{m}_{s'})} (\Delta_{s'p'}^{00} + \Delta_{s'p'}^{01}) \right\},$$

and $\mathbf{q}_1 = (q_{11}, \dots, q_{S\tilde{N}_s})$ is a \tilde{N} -dimensional vector where

$$q_{1,s,p} = -\chi_{1,1-\alpha}^2 \cdot \left(\frac{\tilde{n}_s}{N} \right)^2 \left\{ \frac{\Delta_{sp}^{10} + \Delta_{sp}^{11}}{\tilde{m}_s(\tilde{m}_s - 1)} - \frac{(\Delta_{sp}^{10} + \Delta_{sp}^{11})^2}{\tilde{m}_s^2(\tilde{m}_s - 1)} \right. \\ \left. + \frac{\Delta_{sp}^{00} + \Delta_{sp}^{01}}{(\tilde{n}_s - \tilde{m}_s)(\tilde{n}_s - \tilde{m}_s - 1)} - \frac{(\Delta_{sp}^{00} + \Delta_{sp}^{01})^2}{(\tilde{n}_s - \tilde{m}_s)^2(\tilde{n}_s - \tilde{m}_s - 1)} \right\}.$$

- linear constraints: $\mathbf{q}_{2s}^T \mathbf{x} = P_s$, where \mathbf{q}_{2s} is the zero-one indicator vector for all the \tilde{N}_s possible unique $2 \times 2 \times 2$ tables of $\Lambda_{[s]}$.
- box constraints: $d_{sp} \geq 0$ for all s and p .
- integrality constraints: all \tilde{N} elements of \mathbf{x} are integers.

Appendix C: More Details About the Calculations of Sensitivity Weights and Sensitive Sets

We here only consider Fisher's sharp null hypothesis (i.e., integer programs (P1) and (P2)) as the same method can be applied for Neyman's weak null hypothesis also. For type I randomization designs, let $\tilde{\mathbf{Y}} = (\tilde{Y}_1^{00}, \tilde{Y}_1^{01}, \tilde{Y}_1^{10}, \tilde{Y}_1^{11}, \dots, \tilde{Y}_I^{00}, \tilde{Y}_I^{01}, \tilde{Y}_I^{10}, \tilde{Y}_I^{11})$ be an optimal solution to (P1). According to Definition 3, we have

$$\begin{aligned} W_T^{FP} &= \frac{|\{ij : Z_{ij} = 1, Y_{ij}^* = 1, \tilde{Y}_{ij} = 0\}|}{|\{ij : Y_{ij}^* \neq \tilde{Y}_{ij}\}|} = \frac{\sum_{i=1}^I (\Lambda_i^{11} - \tilde{Y}_i^{11})}{\sum_{i=1}^I (\tilde{Y}_i^{00} + \Lambda_i^{01} - \tilde{Y}_i^{01} + \tilde{Y}_i^{10} + \Lambda_i^{11} - \tilde{Y}_i^{11})}, \\ W_T^{FN} &= \frac{|\{ij : Z_{ij} = 1, Y_{ij}^* = 0, \tilde{Y}_{ij} = 1\}|}{|\{ij : Y_{ij}^* \neq \tilde{Y}_{ij}\}|} = \frac{\sum_{i=1}^I \tilde{Y}_i^{10}}{\sum_{i=1}^I (\tilde{Y}_i^{00} + \Lambda_i^{01} - \tilde{Y}_i^{01} + \tilde{Y}_i^{10} + \Lambda_i^{11} - \tilde{Y}_i^{11})}, \\ W_C^{FP} &= \frac{|\{ij : Z_{ij} = 0, Y_{ij}^* = 1, \tilde{Y}_{ij} = 0\}|}{|\{ij : Y_{ij}^* \neq \tilde{Y}_{ij}\}|} = \frac{\sum_{i=1}^I (\Lambda_i^{01} - \tilde{Y}_i^{01})}{\sum_{i=1}^I (\tilde{Y}_i^{00} + \Lambda_i^{01} - \tilde{Y}_i^{01} + \tilde{Y}_i^{10} + \Lambda_i^{11} - \tilde{Y}_i^{11})}, \\ W_C^{FN} &= \frac{|\{ij : Z_{ij} = 0, Y_{ij}^* = 0, \tilde{Y}_{ij} = 1\}|}{|\{ij : Y_{ij}^* \neq \tilde{Y}_{ij}\}|} = \frac{\sum_{i=1}^I \tilde{Y}_i^{00}}{\sum_{i=1}^I (\tilde{Y}_i^{00} + \Lambda_i^{01} - \tilde{Y}_i^{01} + \tilde{Y}_i^{10} + \Lambda_i^{11} - \tilde{Y}_i^{11})}. \end{aligned}$$

Note that after reformulating (P0) as (P1) or (P2), it is more natural to directly calculate a union of various sensitive sets. Specifically, let \mathcal{S} be a sensitive set given from (P0), G the permutation group over \mathcal{I} defined in Section 3.4.1 and $g\mathcal{S} = \{g(ij) : \tilde{Y}_{ij} \neq Y_{ij}^*\}$ for $g \in G$, then we have

$$\bigcup_{g \in G} g\mathcal{S} = \bigcup_{i=1}^I \{A_i^{00} \cup A_i^{01} \cup A_i^{10} \cup A_i^{11}\},$$

where $A_i^{00} = \{ij : Z_{ij} = 0, Y_{ij}^* = 0, \tilde{Y}_i^{00} \neq 0, j = 1, \dots, n_i\}$, $A_i^{01} = \{ij : Z_{ij} = 0, Y_{ij}^* = 1, \tilde{Y}_i^{01} \neq \Lambda_i^{01}, j = 1, \dots, n_i\}$, $A_i^{10} = \{ij : Z_{ij} = 1, Y_{ij}^* = 0, \tilde{Y}_i^{10} \neq 0, j = 1, \dots, n_i\}$, and

$$A_i^{11} = \{ij : Z_{ij} = 1, Y_{ij}^* = 1, \tilde{Y}_i^{11} \neq \Lambda_i^{11}, j = 1, \dots, n_i\}.$$

For type II randomization designs, let $\tilde{\mathbf{d}} = (\tilde{d}_{11}, \dots, \tilde{d}_{S\tilde{N}_s})$ be an optimal solution to (P2). According to Definition 3, we have

$$\begin{aligned} W_T^{FP} &= \frac{|\{ij : Z_{ij} = 1, Y_{ij}^* = 1, \tilde{Y}_{ij} = 0\}|}{|\{ij : Y_{ij}^* \neq \tilde{Y}_{ij}\}|} = \frac{\sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} \tilde{d}_{sp} (\Lambda_{[s]}^{11} - \Delta_{sp}^{11})}{\sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} \tilde{d}_{sp} (\Delta_{sp}^{00} + \Delta_{sp}^{10} + \Lambda_{[s]}^{01} - \Delta_{sp}^{01} + \Lambda_{[s]}^{11} - \Delta_{sp}^{11})}, \\ W_T^{FN} &= \frac{|\{ij : Z_{ij} = 1, Y_{ij}^* = 0, \tilde{Y}_{ij} = 1\}|}{|\{ij : Y_{ij}^* \neq \tilde{Y}_{ij}\}|} = \frac{\sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} \tilde{d}_{sp} \Delta_{sp}^{10}}{\sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} \tilde{d}_{sp} (\Delta_{sp}^{00} + \Delta_{sp}^{10} + \Lambda_{[s]}^{01} - \Delta_{sp}^{01} + \Lambda_{[s]}^{11} - \Delta_{sp}^{11})}, \\ W_C^{FP} &= \frac{|\{ij : Z_{ij} = 0, Y_{ij}^* = 1, \tilde{Y}_{ij} = 0\}|}{|\{ij : Y_{ij}^* \neq \tilde{Y}_{ij}\}|} = \frac{\sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} \tilde{d}_{sp} (\Lambda_{[s]}^{01} - \Delta_{sp}^{01})}{\sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} \tilde{d}_{sp} (\Delta_{sp}^{00} + \Delta_{sp}^{10} + \Lambda_{[s]}^{01} - \Delta_{sp}^{01} + \Lambda_{[s]}^{11} - \Delta_{sp}^{11})}, \\ W_C^{FN} &= \frac{|\{ij : Z_{ij} = 0, Y_{ij}^* = 0, \tilde{Y}_{ij} = 1\}|}{|\{ij : Y_{ij}^* \neq \tilde{Y}_{ij}\}|} = \frac{\sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} \tilde{d}_{sp} \Delta_{sp}^{00}}{\sum_{s=1}^S \sum_{p=1}^{\tilde{N}_s} \tilde{d}_{sp} (\Delta_{sp}^{00} + \Delta_{sp}^{10} + \Lambda_{[s]}^{01} - \Delta_{sp}^{01} + \Lambda_{[s]}^{11} - \Delta_{sp}^{11})}. \end{aligned}$$

Meanwhile, we can get a collection of sensitive sets

$$\bigcup_{g \in G} g\mathcal{S} = \bigcup_{s=1}^S \bigcup_{p=1}^{\tilde{N}_s} \{B_{sp}^{00} \cup B_{sp}^{01} \cup B_{sp}^{10} \cup B_{sp}^{11}\},$$

where $B_{sp}^{00} = \{ij : Z_{ij} = 0, Y_{ij}^* = 0, \Lambda_i = \Lambda_{[s]}, \tilde{d}_{sp} \neq 0, \Delta_{sp}^{00} \neq 0\}$, $B_{sp}^{01} = \{ij : Z_{ij} = 0, Y_{ij}^* = 1, \Lambda_i = \Lambda_{[s]}, \tilde{d}_{sp} \neq 0, \Delta_{sp}^{01} \neq \Lambda_{[s]}^{01}\}$, $B_{sp}^{10} = \{ij : Z_{ij} = 1, Y_{ij}^* = 0, \Lambda_i = \Lambda_{[s]}, \tilde{d}_{sp} \neq 0, \Delta_{sp}^{10} \neq 0\}$, and $B_{sp}^{11} = \{ij : Z_{ij} = 1, Y_{ij}^* = 1, \Lambda_i = \Lambda_{[s]}, \tilde{d}_{sp} \neq 0, \Delta_{sp}^{11} \neq \Lambda_{[s]}^{11}\}$.

Appendix D: Simulation Studies for Computing Warning Accuracy and Sensitivity Weights with Neyman’s Weak Null

In the main text, we conducted simulation studies on the computational efficiency of the adaptive reformulation strategy for calculating warning accuracy and sensitivity weights proposed in Section 3.4.2 with Fisher’s sharp null. We also obtained some insights on how warning accuracy and sensitivity weights vary with the effect size of measured outcomes and sample size. In this section, we conduct parallel simulation studies with Neyman’s weak null. We investigate both type I and type II randomization designs through considering Simulation Scenario 1 proposed in Section 3.4.3 and Simulation Scenario 2 described as below. As emphasized in the main text, all the data generating processes described in the simulation studies in this paper are only for automatically generating data sets for simulations as our framework works for any given data sets and does not depend on any data generating models.

- **Simulation Scenario 3 (for Type II randomization designs):** We consider a stratified randomized experiment or a stratified observational study (with most strata being small) with $I = 400$ or 2000 strata. We let $\mathcal{U}(A)$ denote the uniform distribution over the set A . In each independent simulation run, for each $i = 1, \dots, I$ we randomly draw m_i from $\mathcal{U}(\{2, 3\})$ and then randomly draw $n_i - m_i$ from $\mathcal{U}(\{2, 3\})$. Then we have $E(N) = \{E(m_i) + E(n_i - m_i)\} \cdot I = 5I = 2000$ or $10,000$.

In each independent simulation run, after generating m_i and $n_i - m_i$ for each stra-

tum i , we follow the same procedure as described in Section 3.4.3 to generate the treatment indicators \mathbf{Z} and the measured outcomes \mathbf{Y}^* based on the prespecified measured effect size (p_0, p_1) . We here consider testing Neyman’s weak null H_0^{weak} . After conducting 1000 independent simulation runs for each of the 18 different prespecified sets of $(E(N), p_0, p_1)$ under Simulation Scenarios 1 and 3, we give the simulations results of the corresponding average computation time, average warning accuracy and average sensitivity weights in Table 3.5. We here report some related details about the specific procedure of obtaining the results in Table 3.5: (i) As mentioned in the main text, conducting a sensitivity analysis or a validation study is typically more meaningful when we detected a treatment effect in a primary analysis (Rosenbaum, 2002b, 2010) based on measured outcomes \mathbf{Y}^* . Therefore, we here exclude few simulation runs (i.e., generated data sets) in which Neyman’s weak null failed to be rejected based on \mathbf{Y}^* (20 out of 36,000 runs). (ii) For the remaining 35,980 simulation runs, to prevent our simulation studies from failing to be finished in a tolerable amount of time, we force a simulation run to stop if it runs more than 100 seconds, report the total number of such cases, and exclude such cases when calculating the average computation time, warning accuracy, and sensitivity weights. However, such potentially computationally infeasible cases (computation time more than 100 seconds) are very rare (17 out of 35,980 runs) and in most cases, our framework is computationally efficient with Neyman’s weak null. (iii) As in Section 3.4.3, all the computation tasks in this section were also done by the optimization solver Gurobi (version 9.1) (Gurobi Optimization, LLC, 2022) and a laptop computer with a 1.6 GHz Dual-Core Intel Core i5 processor and 4 GB 1600 MHz DDR3 memory. From Table 3.5, we can see that the general patterns observed and the insights obtained from Fisher’s sharp null case considered in Table 3.2 (see detailed descriptions in Section 3.4.3) also hold for

Neyman's weak null case.

Table 3.5: Simulations with Neyman's weak null. We report the average computation time (in seconds), warning accuracy \mathcal{WA} and sensitivity weights $(W_T^{FP}, W_T^{FN}, W_C^{FP}, W_C^{FN})$ of different sets of $(E(N), p_0, p_1)$ for Simulation Scenarios 1 and 3 (for type I and type II randomization designs respectively).

| Type I Randomization Designs (Simulation Scenario 1) | | | | | | | | | | | | |
|---|---------------|----------------|------------|------------|------------|------------|-----------------|----------------|------------|------------|------------|------------|
| $p_0 = 0.3$ | $E(N) = 2000$ | | | | | | $E(N) = 10,000$ | | | | | |
| | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.4$ | 0.20 s | 0.98 | 0.32 | 0.00 | 0.00 | 0.68 | 5.62 s | 0.98 | 0.35 | 0.00 | 0.00 | 0.65 |
| $p_1 = 0.6$ | 0.26 s | 0.92 | 0.46 | 0.00 | 0.00 | 0.54 | 6.35 s | 0.91 | 0.46 | 0.00 | 0.00 | 0.54 |
| $p_1 = 0.8$ | 0.35 s | 0.83 | 0.54 | 0.00 | 0.00 | 0.46 | 7.59 s | 0.83 | 0.54 | 0.00 | 0.00 | 0.46 |
| $p_0 = 0.6$ | $E(N) = 2000$ | | | | | | $E(N) = 10,000$ | | | | | |
| | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.7$ | 0.20 s | 0.98 | 0.66 | 0.00 | 0.00 | 0.34 | 5.65 s | 0.98 | 0.65 | 0.00 | 0.00 | 0.35 |
| $p_1 = 0.8$ | 0.25 s | 0.95 | 0.70 | 0.00 | 0.00 | 0.30 | 6.09 s | 0.94 | 0.69 | 0.00 | 0.00 | 0.31 |
| $p_1 = 0.9$ | 0.29 s | 0.91 | 0.74 | 0.00 | 0.00 | 0.26 | 7.15 s | 0.91 | 0.71 | 0.00 | 0.00 | 0.29 |
| $p_0 = 0.9$ | $E(N) = 2000$ | | | | | | $E(N) = 10,000$ | | | | | |
| | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.2$ | 0.45 s | 0.75 | 0.00 | 0.46 | 0.54 | 0.00 | 7.83 s | 0.74 | 0.00 | 0.46 | 0.54 | 0.00 |
| $p_1 = 0.4$ | 0.36 s | 0.83 | 0.00 | 0.37 | 0.63 | 0.00 | 8.73 s | 0.83 | 0.00 | 0.39 | 0.61 | 0.00 |
| $p_1 = 0.6$ | 0.29 s | 0.91 | 0.00 | 0.26 | 0.74 | 0.00 | 7.34 s | 0.91 | 0.00 | 0.29 | 0.71 | 0.00 |
| Type II Randomization Designs (Simulation Scenario 3) | | | | | | | | | | | | |
| $p_0 = 0.3$ | $E(N) = 2000$ | | | | | | $E(N) = 10,000$ | | | | | |
| | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.4$ | 0.63 s | 0.98 | 0.25 | 0.00 | 0.00 | 0.75 | 5.12 s | 0.97 | 0.24 | 0.00 | 0.00 | 0.76 |
| $p_1 = 0.6$ | 0.63 s | 0.90 | 0.45 | 0.00 | 0.00 | 0.55 | 5.11 s | 0.89 | 0.46 | 0.00 | 0.00 | 0.54 |
| $p_1 = 0.8$ | 0.52 s | 0.81 | 0.53 | 0.00 | 0.00 | 0.47 | 4.94 s | 0.80 | 0.53 | 0.00 | 0.00 | 0.47 |
| $p_0 = 0.6$ | $E(N) = 2000$ | | | | | | $E(N) = 10,000$ | | | | | |
| | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | \mathcal{WA} | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
| $p_1 = 0.7$ | 0.62 s | 0.98 | 0.74 | 0.00 | 0.00 | 0.26 | 5.14 s | 0.97 | 0.75 | 0.00 | 0.00 | 0.25 |
| $p_1 = 0.8$ | 0.55 s | 0.94 | 0.68 | 0.00 | 0.00 | 0.32 | 4.95 s | 0.93 | 0.67 | 0.00 | 0.00 | 0.33 |
| $p_1 = 0.9$ | 0.43 s | 0.89 | 0.71 | 0.00 | 0.00 | 0.29 | 4.82 s | 0.89 | 0.70 | 0.00 | 0.00 | 0.30 |
| $p_0 = 0.9$ | $E(N) = 2000$ | | | | | | $E(N) = 10,000$ | | | | | |

| | Time | $\mathcal{W}\mathcal{A}$ | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} | Time | $\mathcal{W}\mathcal{A}$ | W_T^{FP} | W_T^{FN} | W_C^{FP} | W_C^{FN} |
|-------------|--------|--------------------------|------------|------------|------------|------------|--------|--------------------------|------------|------------|------------|------------|
| $p_1 = 0.2$ | 0.56 s | 0.71 | 0.00 | 0.46 | 0.54 | 0.00 | 5.03 s | 0.70 | 0.00 | 0.45 | 0.55 | 0.00 |
| $p_1 = 0.4$ | 0.44 s | 0.81 | 0.00 | 0.40 | 0.60 | 0.00 | 4.84 s | 0.80 | 0.00 | 0.39 | 0.61 | 0.00 |
| $p_1 = 0.6$ | 0.48 s | 0.89 | 0.00 | 0.29 | 0.71 | 0.00 | 4.85 s | 0.89 | 0.00 | 0.29 | 0.71 | 0.00 |

4. Relationship Between Changing Malaria Burden and Low Birth Weight in Sub-Saharan Africa: A Difference-in-Differences Study via A Pair-of-Pairs Approach

This chapter is adapted from “Heng, S., O’Meara, W. P., Simmons, R. A., and Small, D. S. (2021). Relationship between changing malaria burden and low birth weight in sub-Saharan Africa: a difference-in-differences study via a pair-of-pairs approach. *eLife*, 10:e65133.”

4.1 Introduction

In 2018, according to the *World Malaria Report 2019* published by the WHO, an estimated 228 million malaria cases occurred worldwide, with an estimated 405,000 deaths from malaria globally (WHO, 2019). Dellicour et al. (2010) estimated that around 85 million pregnancies occurred in 2007 in areas with stable *Plasmodium falciparum* (one of the most prevalent malaria parasites) transmission and therefore were at risk of malaria. Pregnant women are particularly susceptible to malaria, even if they have developed immunity from childhood infections, in part because parasitized cells in the placenta express unique variant surface antigens (Rogerson

et al., 2007). Women who are infected during pregnancy may or may not experience symptoms, but the presence of parasites has grave consequences for both mother and unborn baby. Parasites exacerbate maternal anemia and they also sequester in the placenta, leading to intrauterine growth restriction, low birth weight (i.e., birth weight < 2,500 grams), preterm delivery and even stillbirth and neonatal death. Preventing malaria during pregnancy with drugs or insecticide treated nets has a significant impact on pregnancy outcomes (Eisele et al., 2012; Kayentao et al., 2013; Radeva-Petrova et al., 2014).

Observational and interventional studies of malaria in pregnant women are complicated by the difficulty of enrolling women early in their pregnancy. However, in one study, early exposure to *Plasmodium falciparum* (before 120 days gestation), prior to initiating malaria prevention measures, was associated with a reduction in birth weight of more than 200 grams and reduced average gestational age of nearly one week (Schmiegelow et al., 2017). For other representative studies on the negative influence of malaria infection during early pregnancy on birth outcomes, see Menendez et al. (2000), Ross and Smith (2006), Huynh et al. (2011), Valea et al. (2012), Walker et al. (2014), and Huynh et al. (2015). These results suggest the impact of malaria infection on stillbirths, perinatal, and neonatal mortality may be substantial and needs more careful examination (Fowkes et al., 2020; Gething et al., 2020).

In the last few decades, malaria burden has declined in many parts of the world. Although the magnitude of the decline is difficult to estimate precisely, some reports suggest that the global cases of malaria declined by an estimated 41% between 2000 and 2015 (WHO, 2016) and the clinical cases of *Plasmodium falciparum* malaria declined by 40% in Africa between 2000 and 2015 (Bhatt et al., 2015). How-

ever, estimates of changing morbidity and mortality do not account for the effects of malaria in pregnancy. In the context of global reductions in malaria transmission, we expect fewer pregnancies are being exposed to infection and/or exposed less frequently. This should result in a significant reduction in preterm delivery, low birth weight and stillbirths. However, declining transmission will also lead to reductions in maternal immunity to malaria. Maternal immunity is important in mitigating the effects of malaria infection during pregnancy as is evidenced by the reduced impact of malaria exposure on the second, third and subsequent pregnancies. Thus we anticipate a complex relationship between declining exposure and pregnancy outcomes that depends on both current transmission and historical transmission and community-level immunity (Mayor et al., 2015).

Understanding the potential causal effect of a reduction in malaria burden on the low birth weight rate is crucial as low birth weight is strongly associated with poor cognitive and physical development of children (McCormick et al., 1992; Avchen et al., 2001; Guyatt and Snow, 2004). Although we know from previous interventional studies that preventing malaria in pregnancy is associated with higher birth weight (Eisele et al., 2012; Radeva-Petrova et al., 2014), we do not know whether natural changes in malaria transmission intensity are similarly associated with improved birth outcomes. To address this question, we make use of the fact that while the overall prevalence of malaria has declined in sub-Saharan Africa, the decline has been uneven, with some malaria-endemic areas experiencing sharp drops and others experiencing little change. We use this heterogeneity to assess whether reductions in malaria prevalence reduce the proportion of infants born with low birth weight in sub-Saharan African countries. Our approach conducts a difference-in-differences study (Card and Krueger, 2000; Angrist and Pischke,

2008; St.Clair and Cook, 2015) by leveraging recent developments in matching, a nonparametric statistical analysis approach that can make studies more robust to bias that can arise from statistical model misspecification (Rubin, 1973, 1979; Hansen, 2004; Ho et al., 2007).

4.2 Materials and Methods

4.2.1 Overview

In this analysis, we combine two rich data sources: 1) rasters of annual malaria prevalence means (Bhatt et al., 2015) and 2) the Demographic and Health Surveys (DHS) (ICF, 2019), and we marry two powerful statistical analysis methods of adjusting for covariates – difference-in-differences (Card and Krueger, 2000; Abadie, 2005; Athey and Imbens, 2006; Angrist and Pischke, 2008; Dimick and Ryan, 2014; St.Clair and Cook, 2015) and matching (Rubin, 1973, 2006; Rosenbaum, 2002b, 2010; Hansen, 2004; Stuart, 2010; Zubizarreta, 2012; Pimentel et al., 2015). We match geographically proximal DHS clusters that were collected in different time periods (early vs. late) and then identify pairs of early/late clusters that have either maintained high malaria transmission intensity or experienced substantial declines in malaria transmission intensity. We then match pairs of clusters that differ in their malaria transmission intensity (maintained high vs. declined) but are similar in other key characteristics. Once these quadruples (pairs of pairs) have been formed, our analysis moves to the individual births within these clusters. We use multiple imputation to categorize missing children’s birth weight records as either low birth weight or not, relying on the size of the child at birth reported subjectively by the mother and other demographic characteristics of the mother. Finally, we estimate

the effect of declined malaria transmission intensity on the low birth weight rate by looking at the coefficient of the malaria prevalence indicator (low vs. high) contributing to the low birth weight rate in a mixed-effects linear probability model adjusted for covariates that are potential confounding variables, the group indicator (individual being within a cluster with declined vs. maintained high malaria transmission intensity), and the time indicator (late vs. early).

4.2.2 Data resources

The data we use in this work comes from the following three sources:

(1) Rasters of annual malaria prevalence: These image data, constructed by the Malaria Atlas Project (MAP) (Hay and Snow, 2006; MAP, 2020), estimate for sub-Saharan Africa the spatial distribution of the *Plasmodium falciparum* parasite rate (i.e., the proportion of the population that carries asexual blood-stage parasites) in children from 2 to 10 years old ($PfPR_{2-10}$) for each year between 2000 and 2015 (Bhatt et al., 2015). $PfPR_{2-10}$ has been widely used for measuring malaria transmission intensity (Metselaar and Van Thiel, 1959; Smith et al., 2007; Bhatt et al., 2015; WHO, 2019) and we use it in this work. The value in each pixel indicates the estimated annual $PfPR_{2-10}$ (ranging from 0 to 1) with a resolution of 5km by 5km.

(2) Demographic and Health Surveys (DHS): The DHS are nationally-representative household surveys mainly conducted in low- and middle- income countries that contain data with numerous health and sociodemographic indicators (Corsi et al., 2012; ICF, 2019). We used the Integrated Public Use Microdata Series' recoding of the DHS variables (IPUMS-DHS) which makes the DHS variables consistent across different years and surveys (Boyle et al., 2019).

(3) Cluster Global Positioning System (GPS) data set: This data set contains the geographical information (longitude, latitude and the indicator of urban or rural) of each cluster in the DHS data. In order to maintain respondent confidentiality, the DHS program randomly displaces the GPS latitude/longitude positions for all surveys, while ensuring that the positional error of the clusters is at most 10 kilometers (at most 5 kilometers for over 99% of clusters) and all the positions stay within the country and DHS survey region (DHS, 2019).

4.2.3 Data selection procedure

In this article, we set the study period to be the years 2000–2015, and correspondingly, all the results and conclusions obtained in this article are limited to the years 2000–2015. We set the year 2000 as the starting point of the study period for two reasons. First, the year 2000 is the earliest year in which the estimated annual $PfPR_{2-10}$ is published by MAP (MAP, 2020). Second, according to Bhatt et al. (2015), “the year 2000 marked a turning point in multilateral commitment to malaria control in sub-Saharan Africa, catalysed by the Roll Back Malaria initiative and the wider development agenda around the United Nations Millennium Development Goals.” We set the year 2015 as the ending point based on two considerations. First, when we designed our study in the year 2017, the year 2015 was the latest year in which the estimated annual $PfPR_{2-10}$ was available to us. We became aware after starting our outcome analysis that MAP has published some post-2015 estimated annual $PfPR_{2-10}$ data since then, but, following Rubin (2007)’s advice to design observational studies before seeing and analyzing the outcome data, we felt it was best to stick with the design of our original study for this report and consider the additional data in a later report. Second, the year 2015 was set as a target year by a series of international goals on malaria control. For example, the United

Nations Millennium Development Goals set a goal to “halt by 2015 and begin to reverse the incidence of malaria” and “the more ambitious target defined later by the World Health Organization (WHO) of reducing case incidence by 75% relative to 2000 levels.” (WHO, 2008a; Bhatt et al., 2015). It is worth emphasizing that although we set the years 2000–2015 as the study period and did not investigate any post-2015 MAP data because of the above considerations, those published or upcoming post-2015 MAP data should be considered or leveraged for future related research or follow-up studies.

After selecting 2000–2015 as our study period, we take the middle point years 2007 and 2008 as the cut-off and define the years 2000–2007 as the “early years” and the years 2008–2015 as the “late years.” We include all the sub-Saharan countries that satisfy the following two criteria: (1) The rasters of estimated annual $PfPR_{2-10}$ between 2000 and 2015 created by the Malaria Atlas Project include that country. (2) For that country, IPUMS-DHS contains at least one standard DHS between 2000–2007 (“early year”) and at least one standard DHS between 2008–2015 (“late year”), and both surveys include the cluster GPS coordinates. If there is more than one early (late) years for which the above data are all available, we chose the earliest early year (latest late year). This choice was made to maximize the time interval for the decrease of malaria prevalence, if any, to have an effect on the birth weight of infants. For those countries that have at least one standard DHS with available cluster GPS data in the late year (2008–2015), but no available standard DHS or GPS data in the early year (2000–2007), we still include them if they have a standard DHS along with its GPS data for the year 1999 (with a possible overlap into 1998). In this case, we assign MAP annual $PfPR_{2-10}$ estimates from 2000 to the 1999 DHS data. This allows us to include two more countries, Cote d’Ivoire and

Tanzania. The 19 sub-Saharan African countries that meet the above eligibility criteria are listed in Table 4.1.

Table 4.1: The 19 selected sub-Saharan African countries along with their chosen early/late years of malaria prevalence (i.e., estimated parasite rate $PfPR_{2-10}$) and IPUMS-DHS early/late years. Note that some span over two successive years.

| Country | Malaria Prevalence | | IPUMS-DHS | |
|---------------------------|--------------------|-----------|------------|-----------|
| | Early Year | Late Year | Early Year | Late Year |
| Benin | 2001 | 2012 | 2001 | 2011–12 |
| Burkina Faso | 2003 | 2010 | 2003 | 2010 |
| Cameron | 2004 | 2011 | 2004 | 2011 |
| Congo Democratic Republic | 2007 | 2013 | 2007 | 2013–14 |
| Cote d'Ivoire | 2000 | 2012 | 1998–99 | 2011–12 |
| Ethiopia | 2000 | 2010 | 2000 | 2010–11 |
| Ghana | 2003 | 2014 | 2003 | 2014 |
| Guinea | 2005 | 2012 | 2005 | 2012 |
| Kenya | 2003 | 2014 | 2003 | 2014 |
| Malawi | 2000 | 2010 | 2000 | 2010 |
| Mali | 2001 | 2012 | 2001 | 2012–13 |
| Namibia | 2000 | 2013 | 2000 | 2013 |
| Nigeria | 2003 | 2013 | 2003 | 2013 |
| Rwanda | 2005 | 2014 | 2005 | 2014–15 |
| Senegal | 2005 | 2010 | 2005 | 2010–11 |
| Tanzania | 2000 | 2015 | 1999 | 2015–16 |
| Uganda | 2000 | 2011 | 2000–01 | 2011 |
| Zambia | 2007 | 2013 | 2007 | 2013–14 |
| Zimbabwe | 2005 | 2015 | 2005–06 | 2015 |

From Table 4.1, we can see that among the 19 countries, only two countries (Congo

Democratic Republic and Zambia) happen to take the margin year 2007 as the early year and no countries take the margin year 2008 as the late year. This implies that our study is relatively insensitive to our way of defining the early years (2000–2007) and the late years (2008–2015) as most of the selected early years and late years in Table 4.1 do not fall near the margin years 2007 and 2008.

4.2.4 Statistical Analysis

4.2.4.1 Motivation and overview of our approach: difference-in-differences via pair-of-pairs

Our approach to estimating the causal effect of reduced malaria burden on the low birth weight rate is to use a difference-in-differences approach (Card and Krueger, 2000; Abadie, 2005; Athey and Imbens, 2006; Angrist and Pischke, 2008; Dimick and Ryan, 2014; St.Clair and Cook, 2015) combined with matching (Rubin, 1973, 2006; Rosenbaum, 2002b, 2010; Hansen, 2004; Stuart, 2010; Zubizarreta, 2012; Pimentel et al., 2015). In a difference-in-differences approach, units are measured in both an early (before treatment) and late (after treatment) period. Ideally, we would like to observe how the low birth weight rate changes with respect to malaria prevalence within each DHS cluster, so that the DHS clusters themselves could be the units in a difference-in-differences approach. However, this is not feasible because within each country over time the DHS samples different locations (clusters) as the representative data of that country. We use optimal matching (Rosenbaum, 1989, 2010; Hansen and Klopfer, 2006) to pair two DHS clusters, one in the early year and one in the late year as closely as possible, mimicking a single DHS cluster measured twice in two different time periods. After this first-step matching, we define the treated units as the high-low pairs of clusters,

meaning that the early year cluster has high malaria prevalence (i.e., $PfPR_{2-10} > 0.4$) while the late year cluster has low malaria prevalence (i.e., $PfPR_{2-10} < 0.2$), and define the control units as the high-high pairs of clusters, meaning that both the early year and late year clusters have high malaria prevalence (i.e., $PfPR_{2-10} > 0.4$) and the absolute difference between their two values of $PfPR_{2-10}$ (one for the early year and one for the late year) is less than 0.1. The difference-in-differences approach (Card and Krueger, 2000; Angrist and Pischke, 2008; Dimick and Ryan, 2014; St.Clair and Cook, 2015) compares the changes in the low birth weight rate over time for treated units (i.e., high-low pairs of clusters) compared to control units (i.e., high-high pairs of clusters) adjusted for observed covariates. The difference-in-differences approach removes bias from three potential sources (Volpp et al., 2007):

- A difference between treated units and control units that is stable over time cannot be mistaken for an effect of reduced malaria burden because each treated or control unit is compared with itself before and after the time at which reduced malaria burden takes place in the treated units.
- Changes over time in sub-Saharan Africa that affect all treated or control units similarly cannot be mistaken for an effect of reduced malaria burden because changes in low birth weight over time are compared between the treated units and control units.
- Changes in the characteristics (i.e., observed covariates) of the populations (e.g., age of mother at birth) in treated or control units over time cannot be mistaken for an effect of reduced malaria burden as long as those characteristics are measured and adjusted for.

The traditional difference-in-differences approach requires a parallel trend assumption, which states that the path of the outcome (e.g., the low birth weight rate) for the treated unit is parallel to that for the control unit (Card and Krueger, 2000; Angrist and Pischke, 2008; Dimick and Ryan, 2014; St.Clair and Cook, 2015). One way the parallel trend assumption can be violated is if there are events in the late period whose effect on the outcome differs depending on the level of observed covariates and those observed covariates are unbalanced between the treated and control units across time (Shadish et al., 2002). For example, suppose that there are advances in prenatal care in the late year that tend to be available more in urban areas, then the parallel trends assumption could be violated if there are more treated units (i.e., high-low pairs of clusters) in urban areas than control units (i.e., high-high pairs of clusters). To make the parallel trend assumption more likely to hold, instead of conducting a difference-in-differences study simply among all the treated and control units, we use a second-step matching to pair treated units with control units on the observed covariates trajectories (from the early year to the late year) to make the treated units and control units similar in the observed covariates trajectories as they would be under randomization (Rosenbaum, 2002b, 2010; Stuart, 2010), and discard those treated or control units that cannot be paired with similar observed covariates trajectories. For example, by matching on the urban/rural indicator trajectories between the treated and control units, we adjust for the potential source of bias resulting from the possibility that there may be advances in prenatal care in the late year that are available more in urban areas.

Another perspective on how our second-step matching helps to improve a difference-in-differences study is through the survey location sampling variability (Fakhouri et al., 2020). Recall that when constructing representative samples,

the DHS are sampled at different locations (i.e., clusters) across time (ICF, 2019; Boyle et al., 2019). Therefore, if we simply implemented a difference-in-differences approach over all the high-low and high-high pairs of survey clusters and did not use matching to adjust for observed covariates, this survey location sampling variability may generate imbalances (i.e., different trajectories) of observed covariates across the treated and control groups, and therefore may bias the difference-in-differences estimator (Heckman et al., 1997). Imbalances of observed covariates caused by the survey location sampling variability may occur in the following three cases: 1) The survey location sampling variability is affecting the treated and control groups in the opposite direction. Specifically, there is some observed covariate for which the difference between the high-low pairs of sampled clusters tends to be larger (or smaller) than the country's overall difference between the high malaria prevalence regions in the early years and the low malaria prevalence regions in the late years and conversely, the difference in that observed covariate between the high-high pairs of sampled clusters tends to be smaller (or larger) than the country's overall difference between the high malaria prevalence regions in the early years and the high malaria prevalence regions in the late years. 2) The survey location sampling variability is affecting the treated and control groups in the same direction but to different extents. 3) The survey location sampling variability only happened in the treated or control group. Specifically, there is some observed covariate for which the difference between the high-low (or high-high) pairs of sampled clusters tends to differ from the country's overall difference between the high malaria prevalence regions in the early years and the low (or high) malaria prevalence regions in the late years, but this is not the case for the high-high (or high-low) pairs of sampled clusters. Using matching as a nonparametric data preprocessing step in a difference-in-differences study can remove this type of

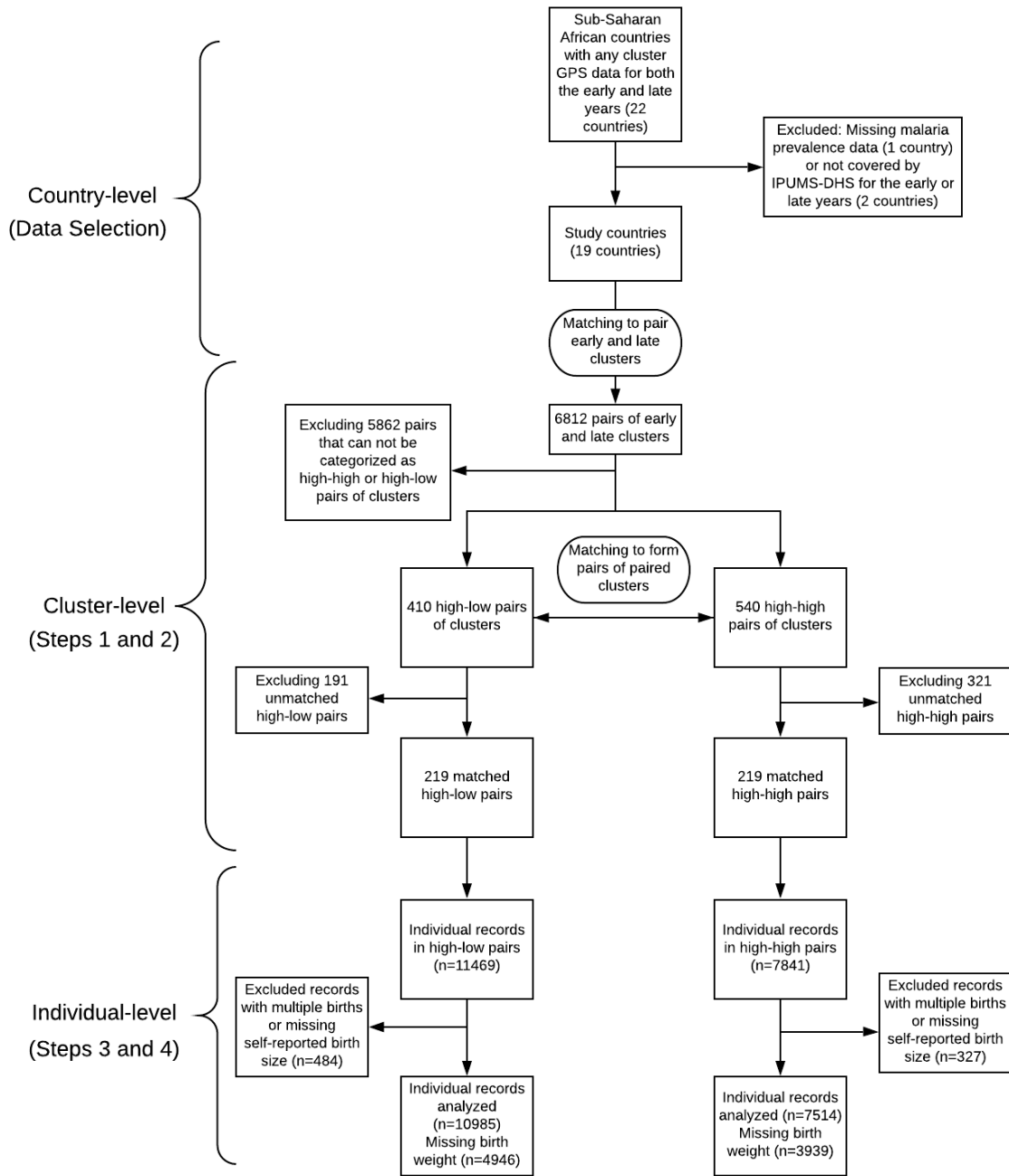
bias because the observed covariates trajectories are forced to be common among the matched treated and control groups (St.Clair and Cook, 2015; Basu and Small, 2020).

An additional important aspect of our approach is that we use multiple imputation to address missingness in the birth weight records. The fraction of missingness in birth weight in the IPUMS-DHS data set is non-negligible and previous studies have noted that failing to carefully and appropriately address the missing data issue with the birth weight records can significantly bias the estimates of the low birth weight rate derived from surveys in developing countries (Boerma et al., 1996; Robles and Goldman, 1999). We address the missing data issue by using multiple imputation with carefully selected covariates. Multiple imputation constructs several plausible imputed data sets and appropriately combines results obtained from each of them to obtain valid inferences under an assumption that the data is missing at random conditional on measured covariates (Rubin, 1987). Our workflow is summarized in Figure 4.1, in which we indicate both the data granularity (country-level, cluster-level, and individual-level) and the corresponding steps of our statistical methodology (including the data selection procedure described in the previous section and the Steps 1–4 of the statistical analysis listed below).

4.2.4.2 Step 1: Proximity prioritized in the matching of high-high and high-low clusters

The DHS collects data from different clusters within the same country in different survey years. To construct pairs of early year and late year clusters which are geographically close such that each pair of clusters can mimic a single cluster measured twice in two different time periods to serve as the unit of a difference-in-

Figure 4.1: Work flow diagram of the study.



differences study, we use optimal matching (Rosenbaum, 1989, 2010; Hansen and Klopfer, 2006) to pair clusters within the same country, one from the early year and one from the late year, based on the geographic proximity of their locations. Specifically, we minimize the total rank-based Mahalanobis distance based on the latitude and longitude of the cluster with a propensity score caliper to pair clusters so that the total distance between the paired early year cluster and late year cluster is as small as possible (Rosenbaum, 1989, 2010; Hansen and Klopfer, 2006). The number of clusters to pair for each country is set to be the minimum of the number of clusters in the early year and the number of clusters in the late year of that country.

4.2.4.3 Step 2: Matching on sociodemographic similarity is emphasized in second matching

We first divide malaria prevalence into three levels with respect to the estimated *Plasmodium falciparum* parasite rates $PfPR_{2-10}$ (ranging from 0 to 1): high ($PfPR_{2-10} > 0.4$), medium ($PfPR_{2-10}$ lies in $[0.2, 0.4]$), and low ($PfPR_{2-10} < 0.2$). For clusters in the year 1999, we use the $PfPR_{2-10}$ in the nearest year in which it is available, i.e., the year 2000. We select the pairs of the early year and late year clusters as formed in Step 1 described above that belong to either one of the following two categories: (1) High-high pairs: both of the estimated parasite rates of the early year and late year clusters within that pair are high (> 0.4), and the absolute difference between the two rates is less than 0.1. (2) High-low pairs: the estimated parasite rate of the early year cluster within that pair is high (> 0.4), while the estimated parasite rate of the late year cluster within that pair is low (< 0.2). 950 out of 6,812 pairs of clusters met one of these two criteria with 540 being high-high pairs and 410 high-low pairs. We removed one high-low pair in

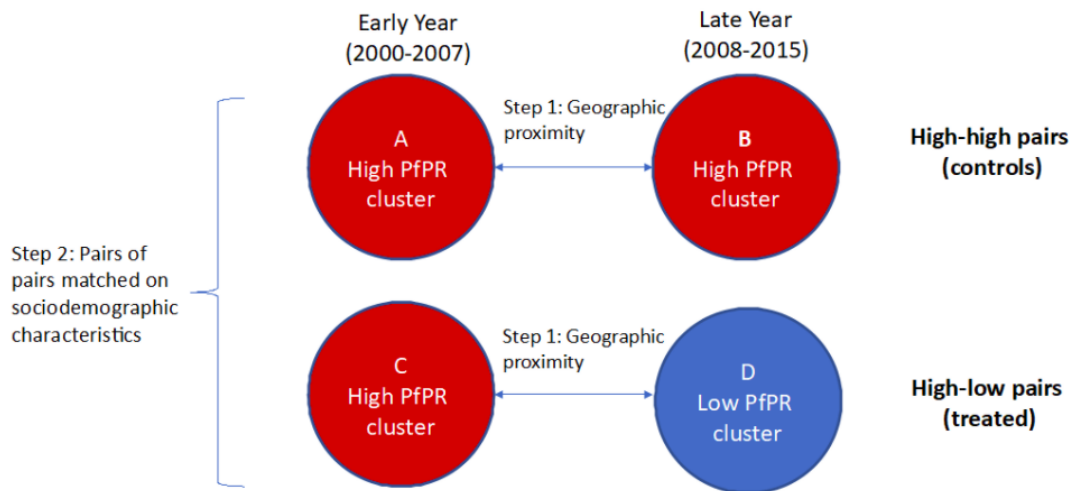
which the late year cluster had an estimated parasite rate value (i.e., $PfPR_{2-10}$) of zero for every year between 2000 and 2015; this cluster was in a high altitude area with temperature unsuitable for malaria transmission and thus was not comparable in malaria transmission intensity to its paired early year cluster with high malaria transmission intensity. Since we would like to study the effect of reduced malaria burden on the low birth weight rate of infants, we consider high-low pairs of clusters as treated units and high-high pairs of clusters as control units, and conduct a matched study by matching each high-low pair with a high-high pair that is similar with respect to covariates that might be correlated with either the treatment (changes in malaria prevalence) or the outcome (low birth weight). We allow matches across different countries. The covariates we match on are cluster averages of the following individual-level covariates, where we code the individual-level covariates as quantitative variables with higher values suggesting higher sociodemographic status:

- Household electricity: 0 – dwelling has no electricity; 1 – otherwise.
- Household main material of floor: 1 – natural or earth-based; 2 – rudimentary; 3 – finished.
- Household toilet facility: 0 – no facility; 1 – with toilet.
- Urban or rural: 0 – rural; 1 – urban.
- Mother’s education level: 0 – no education; 1 – primary; 2 – secondary or higher.
- Indicator of whether the woman is currently using a modern method of contraception: 0 – no; 1 – yes.

The above six sociodemographic covariates were chosen by looking over the variables in the Demographic and Health Surveys (DHS) and choosing those which we thought met the following criteria: 1) The above six covariates are potentially strongly correlated with both the risk of malaria (Baragatti et al., 2009; Krefis et al., 2010; Ayele et al., 2013; Roberts and Matthews, 2016; Sulyok et al., 2017) and birth outcomes (Sahn and Stifel, 2003; Gemperli et al., 2004; Chen et al., 2009; Grace et al., 2015; Padhi et al., 2015), and therefore may be important confounding variables that need to be adjusted for via statistical matching (Rosenbaum and Silber, 2009; Rosenbaum, 2010; Stuart, 2010). 2) The records of the above six covariates are mostly available for all the countries and the survey years in our study samples. Specifically, for the above six covariates, the percentages of missing data (missingness can arise either because the question was not asked or the individual was asked the question but did not respond) among the total individual records from IPUMS-DHS among the 6,812 pairs of clusters remaining after Step 1 are all less than 0.3%.

For each cluster, we define the corresponding six cluster-level covariates by taking the average value for each of the six covariates among the individual records from IPUMS-DHS which are in that cluster, leaving out all missing data. This method of building up cluster-level data from individual-level records from DHS has been commonly used (Kennedy et al., 2011; Larsen et al., 2017). We form quadruples (pairs of pairs) by pairing one high-low pair of clusters (a "treated" unit) with one high-high pair of clusters (a "control" unit), such that all the six cluster-level observed covariates are balanced between both the early and late year clusters for the paired high-low and high-high pairs. We use optimal cardinality matching to form these quadruples (Zubizarreta et al., 2014; Visconti and Zubizarreta, 2018).

Figure 4.2: Formed quadruples (pairs of pairs) of matched high-low and high-high pairs of clusters. In Step 1, pairs of clusters from the early and late time periods are matched on geographic proximity and categorized as ‘high-high’ (comparison, or control) or ‘high-low’ (treated). In Step 2, pairs of high-high clusters are matched with pairs of high-low clusters based on cluster-level sociodemographic characteristics. The difference-in-differences estimate of the coefficient of changing malaria burden on the low birth weight rate is based on comparing $(D-C)$ to $(B-A)$.



Optimal cardinality matching is a flexible matching algorithm which forms the largest number of pairs of treated and control units with the constraint that the absolute standardized differences (absolute value of difference in means in standard deviation units; see [Rosenbaum, 2010](#)) are less than a threshold; we use a threshold of 0.1, which is commonly used to classify a match as adequate ([Neuman et al., 2014](#); [Silber et al., 2016](#)). After implementing the optimal cardinality matching, 219 matched quadruples (pairs of high-low and high-high pairs of clusters) remain. See Figure 4.2 for illustration of the process of forming matched quadruples (pairs of pairs).

4.2.4.4 Step 3: Low birth weight indicator with multiple imputation to address missingness

We then conduct statistical analysis at the individual child level. Among all the 19,310 children's records from the quadruples formed above, we exclude multiple births (i.e., twins, triplets etc), leaving 18,499 records. The outcome variable is the indicator of low birth weight, which is defined as child's birth weight less than 2,500 grams. However, 48% of the birth weight records of children among these 18,499 records are missing. To handle this, we perform multiple imputation, under the assumption of missing at random (Heitjan and Basu, 1996), with 500 replications. An important predictor that is available for imputing the missing low birth weight indicator is the mother's subjective reported size of the child. The mother's reported size of the child is relatively complete in the IPUMS-DHS data set and has been shown to be a powerful tool to handle the missing data problem with birth weight (Blanc and Wardlaw, 2005). We exclude the small number of records with missing mother's subjective reported size of the child, leaving 18,112 records, 47% of which (8,509 records) have missing low birth weight indicator. Among the 9,603 records with observed birth weight, 825 (8.6%) had low birth weight. We first use the `bayesglm` function (part of the `arm` package in R) to fit a Bayesian logistic regression for the outcome of the low birth weight indicator among those children for whom low birth weight is not missing. To make it more plausible that the missing at random assumption holds, the following covariates are included as predictors in this regression because they might affect both missingness and the low birth weight rate:

- The size of the child at birth reported subjectively by the mother: 1 – very small or smaller than average; 2 – average; 3 – larger than average or very

large.

- Mother's age in years.
- Child's birth order number: 1 – the first child born to a mother; 2 – the second, third or fourth child born to a mother; 3 – otherwise.
- Household wealth index: 1 – poorest; 2 – poorer; 3 – middle; 4 – richer; 5 – richest.
- Urban or rural: 0 – rural; 1 – urban.
- Mother's education level: 0 – no education; 1 – primary; 2 – secondary or higher.
- Child's sex: 0 – female; 1 – male.
- Mother's current marital or union status: 0 – never married or formerly in union; 1 – married or living together.
- Indicator of whether the child's mother received any antenatal care while the child was in utero: 0 – no or missing; 1 – yes.

We also include quadratic terms for mother's age in years and child's birth order in the regression since according to [Selvin and Janerich \(1971\)](#), the influences of mother's age and child's birth order on the birth weight do not follow a linear pattern. Note that among the remaining 18,112 records, there are no missing data for all of the above covariates. The prior distributions for the regression coefficients follow the default priors of the `bayesglm` function, i.e., independent Cauchy distributions with center 0 and scale set to 10 for the regression intercept term, 2.5 for binary predictors, and $2.5/(2 \times \text{sd})$ for other numerical predictors, where `sd` is the

standard deviation of the predictor in the data used for fitting the regression (i.e., the 9,603 records with observed birth weight). This default weakly informative prior has been shown to outperform Gaussian and Laplacian priors in a wide variety of settings (Gelman et al., 2008). After fitting this Bayesian logistic regression model, we get the posterior distribution of the regression coefficient associated with each predictor; see Table 4.2. From Table 4.2, we can see that in the imputation model, mother’s age, child’s birth order, mother’s education level, and the mother’s reported birth size are significant predictors, which agrees with the previous literature (e.g., Fraser et al., 1995; Strobino et al., 1995; Richards et al., 2001; de Bernabé et al., 2004).

Table 4.2: Summary of the Bayesian logistic regression model fitted over records with observed birth weight which is used to predict missing low birth weight indicators.

| Predictor | Posterior mean | Posterior std | z-score | p-value |
|--------------------------------------|----------------|---------------|---------|------------|
| (Intercept) | 1.916 | 0.628 | 3.051 | 0.002** |
| Mother’s age (linear term) | −0.207 | 0.045 | −4.562 | < 0.001*** |
| Mother’s age (quadratic term) | 0.003 | 0.001 | 3.987 | < 0.001*** |
| Wealth index | 0.060 | 0.037 | 1.591 | 0.112 |
| Child’s birth order (linear term) | −0.989 | 0.338 | −2.925 | 0.003** |
| Child’s birth order (quadratic term) | 0.211 | 0.086 | 2.447 | 0.014* |
| 0 - rural; 1 - urban | 0.126 | 0.103 | 1.214 | 0.225 |
| Mother’s education level | −0.226 | 0.062 | −3.633 | < 0.001*** |
| Child is boy | −0.068 | 0.083 | −0.815 | 0.415 |
| Mother is married or living together | −0.173 | 0.117 | −1.482 | 0.138 |
| Indicator of antenatal care | −0.046 | 0.093 | −0.493 | 0.622 |
| Indicator of low birth size | 2.410 | 0.090 | 26.776 | < 0.001*** |
| Indicator of large birth size | −1.387 | 0.129 | −10.786 | < 0.001*** |

We then conduct the following procedure in each run of multiple imputation. For each individual with missing birth weight, we first draw from the posterior distribution of the regression coefficients in Table 4.2, we then use these regression coefficients and the individual's covariates (as predictors) to find the probability of the individual having low birth weight and then we use this probability to randomly draw a low birth weight indicator for the individual. We conduct this procedure 500 times, getting 500 independent data sets with imputed low birth weight indicators.

4.2.4.5 Step 4: Estimation of causal effect of reduced malaria burden on the low birth weight rate

For each of the 500 imputed data sets, we then fit a mixed-effects linear probability model where there is a random effect (random intercept) for each cluster to account for the potential correlations between the outcomes among the individual records within the same cluster (Galecki and Burzykowski, 2013). We include in the model the covariates which might be related to both whether an individual is in a high-low vs. high-high pair of clusters and the low birth weight rate. Specifically we include the predictors from the Bayesian logistic regression for multiple imputation as covariate regressors in the mixed-effects linear probability model (listed in Table 4.2), except for the mother's reported birth size. We do not include reported birth size because it is not a pretreatment variable and is a proxy for the outcome (Rosenbaum, 1984). In addition to the above covariates, we include in the model the following three indicators: (1) Low malaria prevalence indicator: indicates whether the individual is from a cluster with a low malaria prevalence ($PfPR_{2-10} < 0.2$). (2) Time indicator: 0 – if the individual is from a early year cluster; 1 – if the individual is from a late year cluster. (3) Group indicator: 0 – if the

individual is from a cluster in a high-high pair of clusters; 1 – if the individual is from a cluster in a high-low pair of clusters. Through adjusting for the time varying covariates via matching and including the above three indicators in the regression, our study uses a difference-in-differences approach for a matched observational study (Wing et al., 2018). Note that even though we do not explicitly incorporate matching into the final model (i.e., the mixed-effects linear probability model (4.1)), matching still reduces the bias due to potential statistical model misspecification in our analysis by being a nonparametric data preprocessing step which makes the distributions of the observed covariates of the selected treated and control units identical or similar, lessening the dependence of the results on the model used to adjust for the observed covariates (Hansen, 2004; Ho et al., 2007). Let $\mathbb{1}(A)$ be the indicator function of event A such that $\mathbb{1}(A) = 1$ if A is true and $\mathbb{1}(A) = 0$ otherwise. To conclude, we consider the following mixed-effects linear probability model for the individual j in cluster i :

$$\begin{aligned} \mathbb{P}(Y_{ij} = 1 \mid i, \mathbf{X}_{ij}) &= k_0 + k_1 \cdot \mathbb{1}(i \text{ is a low malaria prevalence cluster}) \\ &\quad + k_2 \cdot \mathbb{1}(i \text{ is a late year cluster}) \\ &\quad + k_3 \cdot \mathbb{1}(i \text{ is from a high-low pair of clusters}) + \beta^T \mathbf{X}_{ij}, \quad (4.1) \end{aligned}$$

with two error terms $\alpha_i \sim \mathcal{N}(0, \sigma_0)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_1)$.

In Model (4.1), Y_{ij} is the observed outcome (i.e., the low birth weight indicator) and \mathbf{X}_{ij} the covariate regressors (including the quadratic terms of mother’s age and child’s birth order) of the individual j in cluster i , and α_i is the random effect for cluster i . See Table 4.3 for an interpretation of the coefficients of the three indicators and the intercept term (i.e., the k_0, k_1, k_2, k_3) within each matched quadruple. The estimated causal effect of reduced malaria burden (low vs. high malaria

prevalence) on the low birth weight rate is the mean value of the 500 estimated coefficients on the low malaria prevalence indicator obtained (i.e., the k_1) from 500 runs of the mixed-effects linear regression described above. See Appendix 4.5.5 for more details on the statistical inference procedure with multiple imputation, which are also referred to as Rubin’s rules (Carpenter and Kenward, 2012).

Table 4.3: An interpretation of the coefficients of the intercept term and the three indicators defined in model (4.1) (i.e., the k_0, k_1, k_2, k_3) within each matched quadruple. The coefficient of the low malaria prevalence indicator (i.e., the k_1) incorporates the information of the magnitude of the effect of changing malaria burden (from high to low) on the low birth weight rate.

| Cluster | Prevalence | Time | Pair | Coefficients | Within-pair contrast | Between-pair contrast |
|---------|------------|-------|-----------|-------------------------|----------------------|-----------------------|
| 1 | High | Early | High-low | $k_0 + k_3$ | $k_1 + k_2$ | k_1 |
| 2 | Low | Late | High-low | $k_0 + k_1 + k_2 + k_3$ | | |
| 3 | High | Early | High-high | k_0 | k_2 | |
| 4 | High | Late | High-high | $k_0 + k_2$ | | |

It is worth clarifying that although we take a Bayesian approach when imputing (i.e., predicting) the missing low birth weight indicators in Step 3 (i.e., imputation model) and then take a frequentist approach when conducting the 500 separate outcome analyses with the 500 imputed data sets in Step 4 (i.e., substantive model), these two different statistical perspectives (i.e., Bayesian and frequentist) do not conflict with each other when we apply Rubin’s rules to combine these 500 separate outcome analyses as the single estimator and inference reported in Table 4.6. This is because the frequentist validity of applying Rubin’s rules to combine separate outcome analyses with multiple imputed data sets only explicitly depends on the asymptotic normal approximation assumption for each coefficient estimator in Model (4.1) (see Appendix 4.5.5 for more details), and does not directly depend

on how the multiple imputed data sets are generated (e.g., either using a Bayesian imputation model as in Step 3 or using a frequentist imputation model instead). Using a Bayesian imputation model followed by a frequentist substantive model is one of the most common strategies when applying Rubin's rules to conduct statistical inference with multiple imputation; see [Rubin \(1996\)](#), Chapter 3 of [Rubin \(1987\)](#), and Chapter 2 of [Carpenter and Kenward \(2012\)](#). For representative works on justifying the advantages of using a Bayesian imputation model in multiple-imputation inferences, see [Meng \(1994\)](#) and Chapter 2 of [Carpenter and Kenward \(2012\)](#).

4.2.4.6 Secondary analyses

We also conducted the following four secondary analyses (SA1) – (SA4) which examine the causal hypothesis that reduced malaria transmission intensity cause reductions in the low birth weight rate in various ways.

- (SA1) In the first secondary analysis, we fit the mixed-effects linear probability model with multiple imputation only on the children whose age at the corresponding survey is no older than one year old (7,156 out of 18,112 records) to mitigate the potential bias resulting from the births that did not occur in exactly the same year as the year of the corresponding malaria prevalence measurement.
- (SA2) In the second secondary analysis, we fit the mixed-effects linear probability model with multiple imputation over first born children only (3,890 out of 18,112 records) to check if the potential effect of reduced malaria burden on the low birth weight rate is especially substantial/weak for first born children or not.

- (SA3) In the third secondary analysis, we make the difference between high malaria prevalence and low prevalence more extreme. Specifically, we redefine the malaria prevalence levels (ranging from 0 to 1) as: high ($PfPR_{2-10} > 0.45$), medium ($PfPR_{2-10}$ lies in $[0.15, 0.45]$), and low ($PfPR_{2-10} < 0.15$). We then conduct the same statistical analysis procedure as in the primary analysis to check if a moderately greater reduction in malaria burden would lead to more of a decrease in the low birth weight rate or not.
- (SA4) In the fourth secondary analysis, we conduct the same procedure as in (SA3), but making the high-medium-low malaria prevalence cut-offs even more extreme: high ($PfPR_{2-10} > 0.5$), medium ($PfPR_{2-10}$ lies in $[0.1, 0.5]$), and low ($PfPR_{2-10} < 0.1$) to check if a substantially more dramatic reduction in malaria burden would cause a more dramatic decrease in the low birth weight rate or not.

4.2.4.7 Sensitivity analyses

As discussed in the “Motivation and overview of our approach” section, using matching as a data preprocessing step in a difference-in-differences study can reduce the potential bias that may result from a violation of the parallel trend assumption arising from failing to adjust for observed covariates and the survey location sampling variability when using the survey data to conduct a difference-in-differences study. However, neither matching nor difference-in-differences can directly adjust for unobserved covariates (i.e., unmeasured confounders or events). The estimated treatment effect (i.e., the estimated coefficient of the low malaria prevalence indicator contributing to the low birth weight rate) from our primary analysis can be biased by failing to adjust for any potential unobserved covariates.

How potential unobserved covariates may bias the estimated effect in a difference-in-differences study has been understood from various alternative perspectives in the previous literature. These alternative perspectives are intrinsically connected and we briefly list three of them here (for more detailed descriptions, see Appendix 4.5.7):

- Perspective 1: The potential violation of the unconfoundedness assumption (Rosenbaum and Rubin, 1983b; Heckman and Robb, 1985; Heckman et al., 1997).
- Perspective 2: The potential violation of the parallel trend assumption in a difference-in-differences study (Card and Krueger, 2000; Angrist and Pischke, 2008; Hasegawa et al., 2019; Basu and Small, 2020).
- Perspective 3: The difference-in-differences estimator may be biased if there is an event that is more (or less) likely to occur as the intervention happens and the occurrence probability of this event cannot be fully captured by observed covariates (Shadish, 2010; West and Thoemmes, 2010).

To assess the robustness of the results of our primary analysis to potential hidden bias, we adapt an omitted variable sensitivity analysis approach (Rosenbaum and Rubin, 1983a; Imbens, 2003; Ichino et al., 2008; Zhang and Small, 2020). Specifically, our sensitivity analysis model (i.e., Model (4.3) in Appendix 4.5.7) extends Model (4.1) by including a hypothetical unobserved covariate U that is correlated with both the low malaria prevalence indicator and the low birth weight indicator. Specifically, let U_{ij} denote the value of U of individual j in cluster i , we consider

the following data generating process for U_{ij} :

$$\begin{aligned} \mathbb{P}(U_{ij} = 1) = & 50\% + p_1\% \cdot \mathbb{1}(i \text{ is a low malaria prevalence cluster}) \\ & + p_2\% \cdot \mathbb{1}(\text{the observed or the imputed } Y_{ij} = 1), \end{aligned} \quad (4.2)$$

where p_1 and p_2 are prespecified sensitivity parameters of which the unit is a percentage point. Our sensitivity analyses investigate how the estimated treatment effect varies over a range of prespecified values for (p_1, p_2) . See Appendix 4.5.7 for the details of the design of the sensitivity analyses and on how our proposed sensitivity analysis model helps to address the concerns about the hidden bias from Perspectives 1–3 listed above.

4.3 Results

In this section, we report and interpret the results of matching, primary analysis, secondary analyses, and sensitivity analyses relating changes in malaria burden to changes in the birth weight rate between 2000–2015 in sub-Saharan Africa. The R (R Core Team, 2020) code for producing all the main results and tables of this article is posted on GitHub (<https://github.com/siyuheng/Malaria-and-Low-Birth-Weight>).

4.3.1 Matching

We first evaluate the performance of the first-step matching where we focus on the geographical closeness of paired early year and late year clusters from the following three perspectives: (1) the geographic proximity of the early year and the late year clusters within each pair, which is evaluated through the mean distance of

two paired clusters, the within-pair longitude's correlation and latitude's correlation between the paired early year and late year clusters, and the mean values of the longitudes and the latitudes of the paired early year and late year clusters; (2) the closeness of the mean annual malaria prevalence ($PfPR_{2-10}$) of the early year and late year clusters at the early year (i.e., the early malaria prevalence year in Table 4.1); (3) the closeness of the mean annual malaria prevalence of the early year and the late year clusters at the late year (i.e., the late malaria prevalence year in Table 4.1). We report the results in Table 4.4, which indicate that the first step of our matching produced pairs of clusters which are close geographically and in their malaria prevalence at a given time. Of note, the mean Haversine distance of the early year clusters and late year clusters is 24.1 km among the 219 high-low pairs of clusters, and 28.7 km among the 219 high-high pairs of clusters. The within-pair longitudes' and latitudes' correlations between the paired early year and late year clusters among the high-low and high-high pairs are all nearly one.

Table 4.4: The mean Haversine distance of the early year clusters and late year clusters is 24.1 km among the 219 high-low pairs of clusters, and 28.7 km among the 219 high-high pairs of clusters. The within-pair longitudes' and latitudes' correlations between the paired early year and late year clusters among the high-low and high-high pairs all nearly equal one. The mean values of the longitudes, the latitudes, the annual malaria prevalence (i.e., $PfPR_{2-10}$) measured at the early year, denoted as $PfPR_{2-10}^{Early}$, and at the late year, denoted as $PfPR_{2-10}^{Late}$, of the paired early year clusters (clusters sampled at the early year) and late year clusters (clusters sampled at the late year) among the 219 high-low and 219 high-high pairs of clusters used for the statistical inference respectively. Note that an early year cluster has a late year $PfPR_{2-10}$ and a late year cluster has an early year $PfPR_{2-10}$ since the MAP data contain $PfPR_{2-10}$ for each location and for each year between 2000 and 2015.

| | High-low pairs | | High-high pairs | |
|--------------------------------------|----------------|----------|-----------------------|----------------------|
| Mean within-pair Haversine distance | 24.1 km | | 28.7 km | |
| Within-pair correlation of longitude | 0.9999 | | 0.9996 | |
| Within-pair correlation of latitude | 0.9998 | | 0.9997 | |
| | Longitude | Latitude | $PfPR_{2-10}^{Early}$ | $PfPR_{2-10}^{Late}$ |
| Early clusters among high-low pairs | 16.92 | -1.15 | 0.52 | 0.17 |
| Late clusters among high-low pairs | 16.88 | -1.15 | 0.48 | 0.12 |
| Early clusters among high-high pairs | 19.15 | 0.43 | 0.51 | 0.47 |
| Late clusters among high-high pairs | 19.13 | 0.46 | 0.53 | 0.49 |

We then evaluate the performance of the second-step matching, where we focus on the closeness of the sociodemographic status of paired high-low and high-high pairs of clusters, by examining the balance of each covariate among high-low and high-high pairs of early year and late year clusters before and after matching. Recall that for each cluster, we calculate the six cluster-level covariates (i.e., urban or rural, toilet facility, floor facility, electricity, mother's education level, contraception indicator) by averaging over all available individual-level records in that cluster. In each high-low or high-high pair of clusters, there are 12 associated co-

variates, 6 for the early year cluster in that pair and 6 for the late year cluster in that pair. Table 4.5 reports the mean of each covariate among high-low pairs of clusters and high-high pairs of clusters before and after matching, along with the absolute standardized differences before and after matching. From Table 4.5, we can see that before matching, the high-high pairs are quite different from the high-low pairs, all absolute standardized differences are greater than 0.2. The high-low pairs tend to be sociodemographically better off than the high-high pairs (higher prevalence of improved toilet facilities and floor material facilities, higher prevalence of domestic electricity, higher levels of mother's education, higher rate of contraceptive use, and more urban households). To reduce the bias from these observed covariates, we leverage optimal cardinality matching, as described above, to pair a high-low pair of clusters with a high-high pair and throw away the pairs of clusters for which the associated covariates cannot be balanced well. After matching, we can see that all 12 covariates are balanced well – all absolute standardized differences after matching are less than 0.1.

Table 4.5: Balance of each covariate before matching (BM) and after matching (AM). We report the mean of each covariate (including early and late years) for high-low and high-high pairs of clusters, before and after matching. We also report each absolute standardized difference (Std.dif) before and after matching.

| | Before matching | | After matching | | Std.dif | |
|----------------------------|-----------------|-------------|----------------|-------------|---------|------|
| | High-low | High-high | High-low | High-high | BM | AM |
| | (410 pairs) | (540 pairs) | (219 pairs) | (219 pairs) | | |
| Urban/rural (early) | 0.44 | 0.20 | 0.26 | 0.26 | 0.53 | 0.00 |
| Urban/rural (late) | 0.60 | 0.21 | 0.37 | 0.32 | 0.85 | 0.09 |
| Toilet facility (early) | 0.88 | 0.60 | 0.82 | 0.79 | 0.86 | 0.10 |
| Toilet facility (late) | 0.94 | 0.69 | 0.90 | 0.88 | 0.90 | 0.10 |
| Floor material (early) | 1.90 | 1.68 | 1.60 | 1.67 | 0.31 | 0.10 |
| Floor material (late) | 2.22 | 1.79 | 1.92 | 1.87 | 0.59 | 0.07 |
| Electricity (early) | 0.36 | 0.12 | 0.17 | 0.16 | 0.70 | 0.02 |
| Electricity (late) | 0.54 | 0.18 | 0.33 | 0.30 | 0.99 | 0.10 |
| Mother's education (early) | 1.00 | 0.36 | 0.69 | 0.64 | 1.36 | 0.10 |
| Mother's education (late) | 1.23 | 0.42 | 0.87 | 0.83 | 1.78 | 0.10 |
| Contraception (early) | 0.16 | 0.12 | 0.15 | 0.17 | 0.27 | 0.10 |
| Contraception (late) | 0.22 | 0.18 | 0.24 | 0.26 | 0.23 | 0.10 |

4.3.2 Effect of reduced malaria burden on the low birth weight rate

Table 4.9 of Appendix 4.5.1 summarizes the low malaria prevalence indicators, the time indicators, the group indicators, the covariates, and the birth weights of the 18,112 births in the matched clusters. Table 4.6 reports the estimated causal effect of reduced malaria burden (low vs. high malaria prevalence) on the rate of births with low birth weight, which is represented as the coefficient on the malaria

prevalence indicator (diagnostics for the multiple imputation that was used in generating the estimates in Table 4.6 are shown in Table 4.10 of Appendix 4.5.5). We estimate that a decline in malaria prevalence from $PfPR_{2-10} > 0.40$ to less than 0.20 reduces the rate of low birth weight by 1.48 percentage points (95% confidence interval: 3.70 percentage points reduction, 0.74 percentage points increase). A reduction in the low birth weight rate of 1.48 percentage points is substantial; recall that among the study individuals with nonmissing birth weight, the low birth weight rate was 8.6%, so a 1.48 percentage points reduction corresponds to a 17% reduction in the low birth weight rate. The results in Table 4.6 also show that there is strong evidence that mother's age, child's birth order, mother's education level and child's sex are also associated with the low birth weight rate. For example, mothers with higher education level are less likely to deliver a child with low birth weight, and boys are less likely to have low birth weight than girls, which agrees with the previous literature (e.g., Brooke et al., 1989; de Bernabé et al., 2004; Zeka et al., 2008).

Our estimated reduction in the low birth weight rate of 1.48 percentage points from reducing malaria prevalence from high to low is similar to that from a naive difference-in-differences estimator that ignores covariates and missingness of birth weight records. The observed low birth weight rates among the records with observed birth weight within the early year clusters in high-low pairs is 9.33%, in the late year clusters in high-low pairs is 7.52%, in the early year clusters in high-high pairs is 9.18%, and in the late year clusters in high-high pairs is 9.06%. Therefore, the naive difference-in-differences estimator for the effect of reduced malaria burden without adjusting for covariates and missingness of birth weight records is $(7.52\% - 9.33\%) - (9.06\% - 9.18\%) = -1.69\%$ (i.e., 1.69 percentage points reduc-

tion on the low birth weight rate).

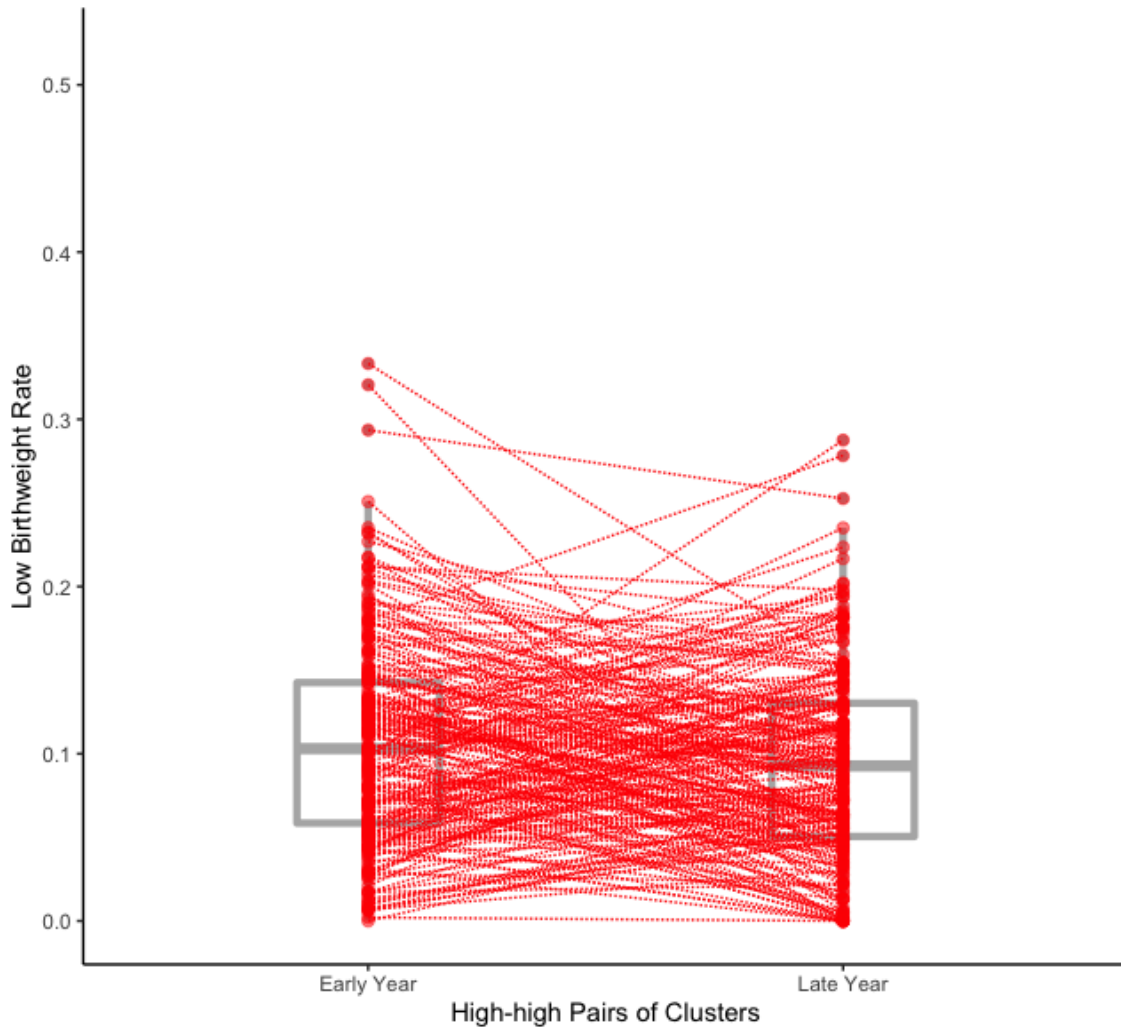
Table 4.6: Inference with multiple imputation and mixed-effects linear probability model (4.1). The unit of estimates and CIs is a percentage point.

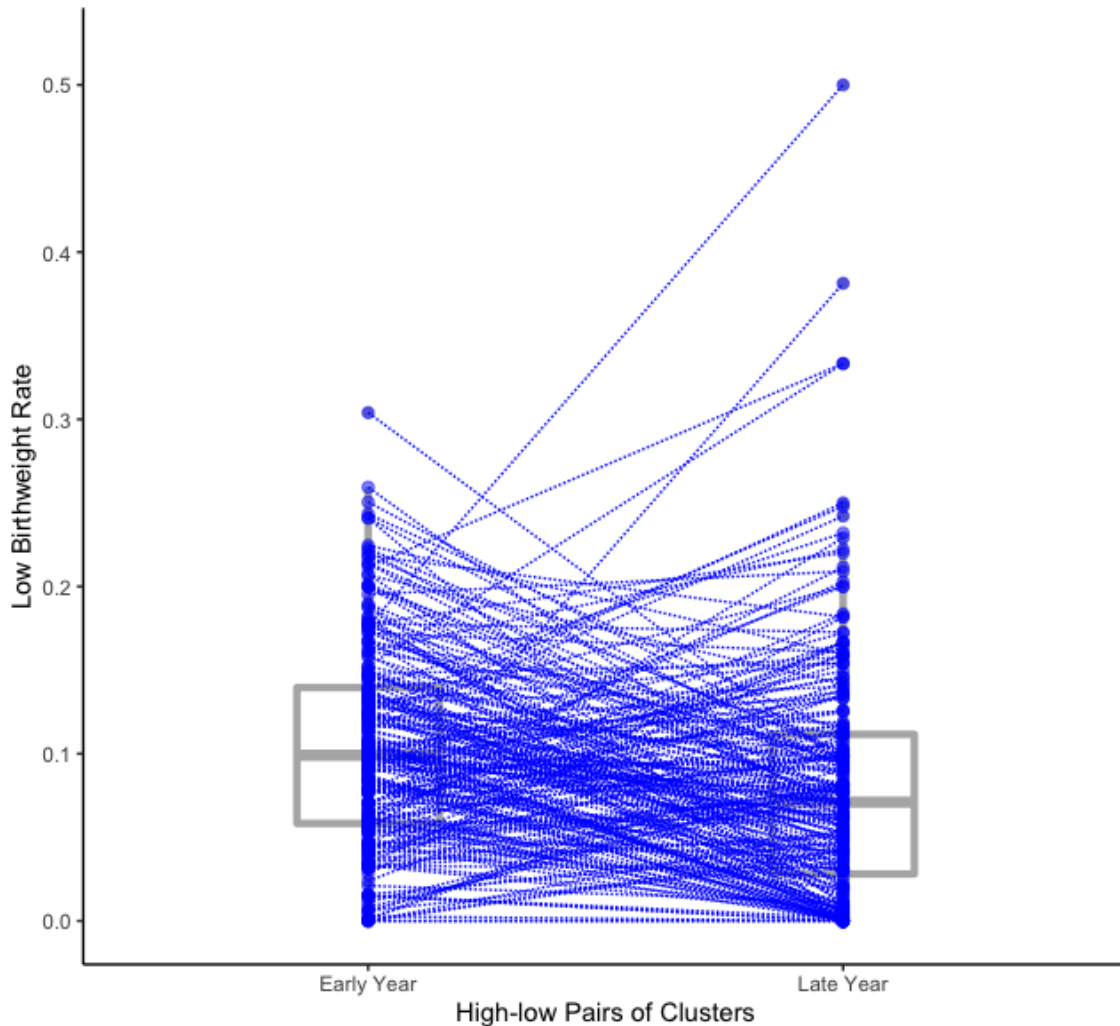
| Regressor | Estimate | 95% CI | p-value |
|---|----------|-----------------|------------|
| 0 - high prevalence; 1 - low prevalence | -1.48 | [-3.70, 0.74] | 0.191 |
| 0 - early year; 1 - late year | -0.06 | [-1.82, 1.69] | 0.943 |
| 0 - high-high pairs; 1 - high-low pairs | 0.21 | [-1.40, 1.82] | 0.797 |
| Mother's age (linear term) | -1.86 | [-2.48, -1.23] | < 0.001*** |
| Mother's age (quadratic term) | 0.03 | [0.02, 0.04] | < 0.001*** |
| Child's birth order (linear term) | -13.91 | [-18.49, -9.32] | < 0.001*** |
| Child's birth order (quadratic term) | 2.91 | [1.82, 4.00] | < 0.001*** |
| Wealth index | 0.09 | [-0.38, 0.56] | 0.709 |
| 0 - rural; 1 - urban | 0.82 | [-0.63, 2.27] | 0.269 |
| Mother's education level | -2.02 | [-2.82, -1.22] | < 0.001*** |
| Child is boy | -1.75 | [-2.75, -0.74] | < 0.001*** |
| Mother is married or living together | -1.43 | [-3.04, 0.19] | 0.083 |
| Antenatal care indicator | -0.96 | [-2.06, 0.13] | 0.085 |

Among all the high-low pairs of clusters in our sample, there has been a decrease in the low birth weight rate from the early years to the late years of 1.81 percentage points (from 9.33% to 7.52%) for records with observed birth weight and an estimated decrease of 2.04 percentage points (from 10.48% to 8.44%) when multiple imputation is used to impute missing birth weight records. We now explore how much of this decrease can be attributed to reduced malaria burden over time. The estimated effect in Table 4.6 of the time indicator (late year vs. early year) is a 0.06 percentage points reduction, which is much less than that of the low malaria

prevalence indicator. Moreover, the estimated change in the low birth weight rate over time among high-low pairs that comes from changes in the covariates over time is a 0.52 percentage points reduction. This is calculated by looking at the difference between $\hat{\beta}^T \bar{x}_{\text{early}}$ and $\hat{\beta}^T \bar{x}_{\text{late}}$, where $\hat{\beta}^T$ is the estimated coefficients of the covariate regressors listed in Table 4.6, and \bar{x}_{early} and \bar{x}_{late} are the average values in high-low pairs of the covariate regressors of the individuals within the early year clusters and those within the late year clusters respectively. These results suggest that after adjusting for the observed covariates listed in Table 4.6 and missingness of birth weight records, the observed decrease in the low birth weight rate over time in high-low pairs comes mainly from reduced malaria burden over time instead of changes over time in the low birth weight rate that affect both high-low and high-high pairs of clusters. To illustrate this point and further verify the potentially substantial effect of reduced malaria burden on the low birth weight rate, we also plot the estimated low birth weight rate of each cluster among the high-high pairs and high-low pairs in our study sample in Figure 4.3. From Figure 4.3, we can see that although in general, for both high-high pairs and high-low pairs, the birth weight rates of the late year clusters are lower than those of the early year clusters, it is clear that the reductions in low birth weight rate from early year to late year among the high-low pairs are considerably greater than those among high-high pairs, suggesting that reducing community-level malaria burden can potentially substantially reduce the low birth weight rate.

Figure 4.3: The estimated low birth weight rate of each cluster within the 219 high-high pairs and the 219 high-low pairs. The estimated low birth weight rate for each cluster are obtained from averaging over all the 500 imputed data sets of the 18,112 individual records. We draw a line to connect two paired clusters (one early year cluster and one late year cluster). Box plots for the low birth weight rates are also shown. Two of the four outliers of the late year clusters among the high-low pairs (i.e., the top four late year clusters in terms of low birth weight rate among the high-low pairs) may result from their extremely small within-cluster sample sizes (no more than 3 individual records for both two clusters).





4.3.3 Results of secondary analyses

The results of our secondary analyses support the interpretation of our primary analysis:

- (SA1) In the first secondary analysis, when only conducting statistical analysis among children whose age at the survey year is no older than 1 year, the point estimate of the coefficient of the low malaria prevalence indicator (1 if $PfPR_{2-10} < 0.2$) is -1.31 percentage points (95% CI: $[-4.70, 2.08]$), which in

general agrees with the result of our primary analysis and implies that our causal conclusion drawn from the primary analysis is relatively robust to the potential hidden bias caused by the births that occurred in different years from the years of the malaria prevalence measurement.

- (SA2) In the second secondary analysis, performing our statistical analysis among first born children only, the estimated coefficient of the low malaria prevalence indicator is -3.73 percentage points (95% CI: $[-9.11, 1.64]$). This implies that the effect of reduced malaria burden on the low birth weight rate may be especially substantial among first born children.
- (SA3) In the third secondary analysis, after slightly enlarging the difference between high malaria prevalence and low prevalence and repeating the two-stage matching procedure described above, there remain 100 high-high pairs of clusters and 100 high-low pairs, with 8,611 individual records remaining in the final model. In (SA3), the point estimate of the coefficient of low malaria prevalence indicator is -1.48 percentage points (95% CI: $[-4.44, 1.48]$). In this case, slightly enlarging the gap between the cutoffs for high/low malaria prevalence did not result in an obvious additional reduction in the low birth weight rate. A possible reason is that the new cut-offs are just slightly different from the previous ones and the changes may still lie within the margin of error of measuring the $PfPR_{2-10}$ or there may not be enough power. In thinking about the results of (SA3), it is useful to also consider the results from (SA4).
- (SA4) In the fourth secondary analysis, after making the high prevalence and low prevalence cut-offs quite extreme and repeating the two-stage matching procedure, there remain 35 high-high pairs of clusters and 35 high-low pairs,

with 3,135 individual records remaining in the final model. In (SA4), the point estimate of the coefficient of low malaria prevalence indicator is -3.04 percentage points (95% CI: $[-8.50, 2.41]$). This implies that a more dramatic reduction in malaria burden can potentially lead to a more dramatic decrease in the low birth weight rate and supports the above hypothesis that the fact that slightly enlarging the gap between the high/low malaria prevalence cut-offs in (SA3) did not result in an evident additional reduction in the low birth weight rate may be due to the potential measurement error of the $PfPR_{2-10}$ or lack of power.

4.3.4 Results of the sensitivity analyses

Recall that in the “Sensitivity analyses” section and Appendix 4.5.7, our sensitivity analyses consider a hypothetical unobserved covariate U that is correlated with both the low malaria prevalence indicator and the low birth weight indicator. For various values of the sensitivity parameters (p_1, p_2) , we report the corresponding point estimates and 95% CIs of the estimated treatment effect (i.e., the coefficient of the low malaria prevalence indicator contributing to the low birth weight rate) in Table 4.11 of Appendix 4.5.7. The results from Table 4.11 of Appendix 4.5.7 show that the estimated treatment effect ranges from 1.13 percentage points reduction to 1.83 percentage points reduction (on the low birth weight rate) if both p_1 and p_2 are between -10 and 10 . Recall that p_1 (or p_2) equals 10 (or -10) means that the probability of the U taking value 1 increases (or decreases) by 10 percentage points if the individual’s low malaria prevalence indicator (or the low birth weight rate indicator) equals 1. That is, allowing both the magnitude of p_1 and the magnitude of p_2 can be up to 10 means that we allow the existence of a nontrivial magnitude of unmeasured confounding in our sensitivity analyses. Therefore, the estimated

treatment effect ranging from 1.13 percentage points reduction to 1.83 percentage points reduction when both p_1 and p_2 are between -10 and 10 means that the magnitude of the estimated treatment effect is still evident (no less than 1.13 percentage points) even if the magnitude of unmeasured confounding is nontrivial (both $|p_1|$ and $|p_2|$ can be up to 10). See Appendix 4.5.7 for the detailed results and interpretations of the sensitivity analyses.

To conclude, although the confidence intervals of the coefficient of the low malaria prevalence indicator on the low birth weight rate presented in the “Results” section cannot exclude a possibility of no effect at level 95% based on our proposed study sample selection procedure and statistical approach, the results and the corresponding interpretations of the primary analysis, the secondary analyses, and the sensitivity analyses have contributed to the weight of the evidence that reduced malaria burden has an important influence on the low birth weight rate in sub-Saharan Africa at the population level.

4.4 Discussion

We have developed a pair-of-pairs matching approach to conduct a difference-in-differences study to examine the causal effect of a reduction in malaria prevalence on the low birth weight rate in sub-Saharan Africa during the years 2000–2015. Although we cannot rule out no effect at a 95% confidence level, the magnitude of the estimated effect of a reduction from high malaria prevalence to low malaria prevalence on the low birth weight rate (1.48 percentage points) is even greater than the estimated effect of a factor thought to be important, antenatal care during pregnancy (0.96 percentage points). In a secondary analysis, we find that reduction in malaria burden from high to low is estimated to be especially crucial for reduc-

ing the low birth weight rate of first born children, reducing it by 3.73 percentage points (95% CI: 9.11 percentage points reduction, 1.64 percentage points increase). This agrees with previous studies which demonstrate that the effects of malaria on birth outcomes are most pronounced in the first pregnancy (e.g., [McGregor et al., 1983](#)).

Previous studies have shown that individual malaria prevention during pregnancy reduces the chances of the woman's baby having low birth weight ([Kayentao et al., 2013](#)). In this paper, we examine the community-level effect of reductions in malaria on pregnancy outcomes as opposed to the individual-level effect of malaria prevention interventions during pregnancy. Our results support extrapolation of studies of antenatal malaria interventions on birth weight to populations experiencing declining malaria burden. Furthermore, we conclude that reports of declining malaria mortality underestimate the contribution of reduced malaria exposure during pregnancy on pregnancy outcomes and neonatal survival. Although some studies have documented higher rates of adverse pregnancy outcomes in malaria-infected women with declining antimalarial immunity, such as may be seen in communities with declining malaria exposure ([Mayor et al., 2015](#)), our study demonstrates that overall reduction in exposure to infection, including during pregnancy, outweighs these individual changes in risk once infected.

Strengths of our study include that we use state-of-the-art causal inference methods on a large representative data set. We develop a novel pair-of-pairs matching approach to conduct a difference-in-differences study to estimate the real world effectiveness of public health interventions by combining DHS data with other data sources. There are two major difficulties when using the DHS data to conduct a difference-in-differences study. The first major difficulty is that within each coun-

try the DHS samples different locations (clusters) over different survey years. Our first-step matching handles this difficulty through using optimal matching to pair the early year DHS clusters and the late year DHS clusters within the same country based on the geographic proximity of their locations. Then each formed pair of clusters can mimic a single cluster measured twice in two different survey years, which serves as the foundation of a difference-in-differences study. The second major difficulty is that although an advantage of the DHS data is that they contain many potentially important cluster-level and individual-level covariates, it may be difficult to come up with a statistical model that is both efficient and robust to adjust for these covariates. A traditional approach to estimating the real world effectiveness of an intervention in such settings is to run a regression of an outcome of interest on a measure of adherence to the treatment (zero if in the period before the intervention was available and ranging from 0 to 1 after the intervention was available), covariates (individual-level and cluster-level covariates) and a random effect for the cluster (Goetgeluk and Vansteelandt, 2008). This regression approach relies heavily on correct specification of the model by which the covariates affect the outcome (e.g., linear, quadratic, cubic), therefore the result can be severely biased by model misspecification (Rubin, 1973, 1979; Hansen, 2004; Ho et al., 2007). We instead use a second-step matching to first optimally select and match the treated units (i.e., high-low pairs of clusters) and control units (i.e., high-high pairs of clusters) to ensure that they have balanced distributions of covariates across time and then run the regression with the dummy variables for the matched sets. Such a nonparametric data preprocessing step before running a regression can potentially reduce bias due to model misspecification (Rubin, 1973, 1979; Hansen, 2004; Ho et al., 2007).

Our merged study data set makes use of two aspects of the richness of the relevant data resources. First, from the perspective of sample size and length of time span, the data set includes over 18,000 births in 19 countries in sub-Saharan Africa and describes changes in the low birth weight rate over a 15 year period. Some of the studied regions had substantial changes in malaria parasite prevalence during this time period, whereas others did not, which provides us ample heterogeneity necessary for conducting a difference-in-differences study. Second, from the perspective of the comprehensiveness of information, our merged data set includes various types of information: from cluster-level to individual-level records; from geographic to sociodemographic characteristics; from surveyed data to predicted data.

Some potential limitations of our study should be considered. First, we discretized the mean malaria prevalence (i.e., $PfPR_{2-10}$ from 0 to 1) into high ($PfPR_{2-10} > 0.4$), medium ($PfPR_{2-10}$ lies in $[0.2, 0.4]$), and low ($PfPR_{2-10} < 0.2$), which means that the magnitude of the estimated causal effect depends on how we define these cut-offs. Our primary analysis suggests that reducing the malaria burden from high to low may substantially help control the low birth weight rate, and our secondary analyses suggest that a more dramatic reduction in malaria prevalence can lead to a more dramatic drop in the low birth weight rate. More research needs to be done on the minimum magnitude of the reduction in malaria prevalence that is needed to cause a substantial drop in the low birth weight rate. Second, we assigned the malaria prevalence (i.e., $PfPR_{2-10}$) data to children's records based on the DHS survey years which may not be exactly the same years as children's actual birth years. For example, a child whose age is three years at the corresponding DHS survey year should have been born three years earlier before that DHS survey

year, in which case we might have assigned the wrong $PfPR_{2-10}$ to that child's gestational period. We examined this issue via SA1 and the result suggested that this did not induce much bias to the results of our primary analysis.

The novel design-based causal inference approach developed in this work, a pair-of-pairs matching approach to conduct a difference-in-differences study (i.e., the two-step matching procedure to form matched pairs of pairs as a nonparametric data preprocessing step in a difference-in-differences study), is potentially useful for researchers who would like to reduce the estimation bias due to potential model misspecification in the traditional difference-in-differences approach. Moreover, the general statistical methodology developed in this work can be applied beyond the malaria settings to handle the heterogeneity of survey time points and locations in data sets such as the Demographic and Health Surveys (DHS).

In summary, the contribution of malaria to stillbirth and neonatal mortality, for which low birth weight is a proxy, are currently not accounted for in global estimates of malaria mortality. Using a large representative data set and innovative statistical evidence, we found point estimates that suggested that reductions in malaria burden at the community level substantially reduce the low birth weight rate. To our knowledge, this is the first study of its kind to evaluate the causal effects of malaria control on birth outcomes using a causal inference framework. Although our confidence intervals do include a possibility of no effect, the evidence from our primary analysis and secondary analyses is strong enough to merit further study and motivate further investments in mitigating the intolerable burden of malaria.

4.5 Appendices

4.5.1 More details on the data selection procedure

We give more details on how we select the study countries (among all sub-Saharan African countries) and their corresponding late year and early year for each of the three data sets: malaria prevalence data (MAP data), IPUMS-DHS data, and DHS cluster GPS data. We define “early year” as 2000–2007 and “late year” as 2008–2015. We first select countries that have both IPUMS-DHS data and DHS GPS data for at least one year between 2000–2007 and one year between 2008–2015. If there are more than one early (late) years available, we choose the earliest early year and latest late year. Note that some DHS can span over two years. In this case, we stick to the way how IPUMS-DHS codes the year of that DHS data set. For example, both Malawi and Tanzania have a standard DHS with GPS data that spans over 2015–2016. We call them Malawi 2015-2016 DHS and Tanzania 2015–2016 DHS respectively. In IPUMS-DHS, the year for Malawi 2015–2016 DHS is coded as 2016, and that for Tanzania 2015-2016 DHS is coded as 2015. Therefore, for Malawi, we use Malawi 2010 DHS as the study sample for the late year and exclude Malawi 2015–2016 DHS. While for Tanzania, we use Tanzania 2015 DHS for the late year. As we have mentioned in the main text, if a country has at least one year between 2008–2015 with available IPUMS-DHS data of which the GPS data is also available, but no available IPUMS-DHS data or the corresponding GPS data between 2000–2007, we still include that country if it has IPUMS-DHS data along with the corresponding GPS data for the year 1999 (possibly with overlap into 1998). This selection procedure results in 19 study countries in total. Note that for the DHS that span over two successive years, sometimes IPUMS-DHS and the GPS data code their years in different ways. In these cases, when attaching

the malaria prevalence data to each cluster, we stick to the year which is used by the GPS data; see Table 4.7 of Appendix 4.5.1. For example, for Benin 2011–2012 DHS, IPUMS-DHS codes its year as 2011 while the GPS data codes its year as 2012. In these cases, we use the malaria prevalence data for 2012 for the clusters within Benin 2011–2012; see the row “Benin (BJ)” in Table 4.7 of Appendix 4.5.1.

Table 4.7: The early and late years coded in the IPUMS-DHS and GPS data sets.

| Country | GPS Data | | Malaria Prevalence | | IPUMS-DHS | |
|--------------------------------|----------|------|--------------------|------|-----------|------|
| | Early | Late | Early | Late | Early | Late |
| Benin (BJ) | 2001 | 2012 | 2001 | 2012 | 2001 | 2011 |
| Burkina Faso (BF) | 2003 | 2010 | 2003 | 2010 | 2003 | 2010 |
| Cameron (CM) | 2004 | 2011 | 2004 | 2011 | 2004 | 2011 |
| Congo Democratic Republic (CD) | 2007 | 2013 | 2007 | 2013 | 2007 | 2013 |
| Cote d’Ivoire (CI) | 1998 | 2012 | 2000 | 2012 | 1998 | 2011 |
| Ethiopia (ET) | 2000 | 2010 | 2000 | 2010 | 2000 | 2011 |
| Ghana (GH) | 2003 | 2014 | 2003 | 2014 | 2003 | 2014 |
| Guinea (GN) | 2005 | 2012 | 2005 | 2012 | 2005 | 2012 |
| Kenya (KE) | 2003 | 2014 | 2003 | 2014 | 2003 | 2014 |
| Malawi (MW) | 2000 | 2010 | 2000 | 2010 | 2000 | 2010 |
| Mali (ML) | 2001 | 2012 | 2001 | 2012 | 2001 | 2012 |
| Namibia (NM) | 2000 | 2013 | 2000 | 2013 | 2000 | 2013 |
| Nigeria (NG) | 2003 | 2013 | 2003 | 2013 | 2003 | 2013 |
| Rwanda (RW) | 2005 | 2014 | 2005 | 2014 | 2005 | 2014 |
| Senegal (SN) | 2005 | 2010 | 2005 | 2010 | 2005 | 2010 |
| Tanzania (TZ) | 1999 | 2015 | 2000 | 2015 | 1999 | 2015 |
| Uganda (UG) | 2000 | 2011 | 2000 | 2011 | 2001 | 2011 |
| Zambia (ZM) | 2007 | 2013 | 2007 | 2013 | 2007 | 2013 |
| Zimbabwe (ZW) | 2005 | 2015 | 2005 | 2015 | 2005 | 2015 |

4.5.2 Country summary

Table 4.8: The numbers of the high-high pairs of clusters and high-low pairs of clusters contributed by each of the 19 selected sub-Saharan African countries after the matching in Step 1 and Step 2. We also summarize the total number of pairs of clusters after Step 1 matching in the first column.

| Country | Step 1 matching | | | Step 2 matching | |
|---------------------------|-----------------|-----------|----------|-----------------|----------|
| | Total pairs | High-high | High-low | High-high | High-low |
| Benin | 247 | 29 | 6 | 4 | 6 |
| Burkina Faso | 400 | 150 | 0 | 19 | 0 |
| Cameron | 466 | 17 | 163 | 16 | 51 |
| Congo Democratic Republic | 300 | 11 | 55 | 11 | 24 |
| Cote d'Ivoire | 140 | 19 | 2 | 7 | 2 |
| Ethiopia | 539 | 0 | 0 | 0 | 0 |
| Ghana | 412 | 24 | 18 | 18 | 8 |
| Guinea | 295 | 47 | 12 | 10 | 12 |
| Kenya | 400 | 2 | 10 | 2 | 8 |
| Malawi | 560 | 96 | 15 | 81 | 15 |
| Mali | 402 | 101 | 21 | 17 | 19 |
| Namibia | 260 | 0 | 0 | 0 | 0 |
| Nigeria | 362 | 24 | 11 | 16 | 1 |
| Rwanda | 462 | 0 | 0 | 0 | 0 |
| Senegal | 376 | 0 | 0 | 0 | 0 |
| Tanzania | 176 | 0 | 68 | 0 | 57 |
| Uganda | 298 | 19 | 29 | 17 | 16 |
| Zambia | 319 | 1 | 0 | 1 | 0 |
| Zimbabwe | 398 | 0 | 0 | 0 | 0 |
| Total | 6812 | 540 | 410 | 219 | 219 |

4.5.3 Some remarks on the IPUMS-DHS data used in this article

There are different units of analysis for data browsing in IPUMS-DHS (Boyle et al., 2019). In “Step 2: Matching on sociodemographic similarity is emphasized in second matching,” for the covariates “Household electricity,” “Household main material of floor,” and “Household toilet facility,” the IPUMS-DHS data we used is at the household members level (each record is a household member). For the covariates “Mother’s education level” and “Indicator of whether the woman is currently using a modern method of contraception,” the IPUMS-DHS data we used is at the birth level (each record is a birth reported by a woman of childbearing age). The covariate “Urban or rural” obtained from the DHS GPS data is at the DHS clusters level. In “Step 3: Low birth weight indicator with multiple imputation to address missingness” and “Step 4: Estimation of causal effect of reduced malaria burden on the low birth weight rate,” the IPUMS-DHS data we used is at the child level (each record is a child under age 5).

4.5.4 More details on the final study population

Table 4.9: Summary of the low malaria prevalence indicators, the time indicators, the group indicators, the covariates, and the birth weight records among the 18,112 study individual records.

| Variables | Percentages of some categories |
|----------------------------------|--|
| Low malaria prevalence indicator | High prevalence (70.6%); Low prevalence (29.4%) |
| Time indicator | Early year (50.3%) Late year (49.7%) |
| Group indicator | High-high pairs (40.9%) |

| | |
|-----------------------------|--|
| | High-low pairs (59.1%) |
| Mother's age in years | ≤ 19 (7.1%) 20 – 29 (52.5%) 30 – 39 (31.4%) ≥ 40 (8.9%) |
| Wealth index | Poorest (20.2%) Poorer (23.3%) Middle (22.8%) Richer (20.4%) Richest (13.3%) |
| Child's birth order | 1 (21.5%) 2 – 4 (46.0%) 4+ (32.6%) |
| Urban or rural | Rural (77.1%) Urban (22.9%) |
| Mother's education level | No education (36.6%) Primary (47.2%) Secondary or higher (16.2%) |
| Child's sex | Female (49.3%) Male (50.7%) |
| Mother's marital status | Never married or formerly in union (11.6%) Married or living together (88.4%) |
| Indicator of antenatal care | Yes (61.9%) No or missing (38.1%) |
| Self-reported birth size | Very small or smaller than average (13.0%) |

| | |
|----------------------------|---|
| | Average (45.5%) |
| | Larger than average or very large (41.5%) |
| Low birth weight indicator | Yes (4.6%) |
| | No (48.5%) |
| | Missing (47.0%) |

4.5.5 Statistical inference with multiple imputation applying Rubin's rules

We apply Rubin's rules (Rubin, 1987; Schafer, 1999; Carpenter and Kenward, 2012) to combine all the imputed data sets to obtain the point estimate, the p-value, and the 95% confidence interval for each coefficient in the mixed-effects linear probability model (4.1) summarized in Table 4.6 of the main text. Suppose that there are M imputed data sets ($M = 500$ in our study). Suppose that for the m -th imputed data set, $m = 1, \dots, 500$, the estimate for the coefficient of the i -th regressor γ_i (including the intercept term), $i = 1, \dots, 14$, is $\hat{\gamma}_{m,i}$, and let V_i be its squared standard error and $\hat{V}_{m,i}$ be the estimated squared standard error from the m -th imputed data set. Suppose that the following normal approximations hold

$$(\hat{\gamma}_{m,i} - \gamma_i) / \sqrt{\hat{V}_{m,i}} \sim \mathcal{N}(0, 1), \quad i = 1, \dots, 14, \quad m = 1, \dots, 500.$$

According to Rubin's rules (Rubin, 1987; Schafer, 1999; Carpenter and Kenward, 2012), we estimate γ_i with $\bar{\gamma}_i = M^{-1} \sum_{m=1}^M \hat{\gamma}_{m,i}$. Consider the corresponding between-imputation variance $B_i = (M - 1)^{-1} \sum_{m=1}^M (\hat{\gamma}_{m,i} - \bar{\gamma}_i)^2$ and the within-

imputation variance $\bar{V}_i = M^{-1} \sum_{m=1}^M \hat{V}_{m,i}$. Then the estimated total variance is

$$T_i = (1 + M^{-1})B_i + \bar{V}_i, \quad i = 1, \dots, 14.$$

Then we can get the corresponding two-sided p-values and 95% confidence intervals based on a Student's t -approximation

$$(\bar{\gamma}_i - \gamma_i) / \sqrt{T_i} \sim t_{v_i}, \quad i = 1, \dots, 14,$$

with degrees of freedom

$$v_i = (m - 1) \left[1 + \frac{\bar{V}_i}{(1 + M^{-1})B_i} \right]^2.$$

4.5.6 Multiple imputation diagnostics

Table 4.10: Diagnostics for multiple imputation with the mixed-effects linear probability model. We report the between-imputation variance ("Between var"), the within-imputation variance ("Within var"), and the variance ratio: (between-imputation variance)/(within-imputation variance), denoted as "Var ratio".

| Regressor | Between var | Within var | Var ratio |
|---|------------------------|-----------------------|-----------|
| 0 - high prevalence; 1 - low prevalence | 3.21×10^{-5} | 9.62×10^{-5} | 0.334 |
| 0 - early year; 1 - late year | 2.20×10^{-5} | 5.81×10^{-5} | 0.379 |
| 0 - high-high pairs; 1 - high-low pairs | 1.92×10^{-5} | 4.83×10^{-5} | 0.398 |
| Mother's age (linear term) | 3.32×10^{-6} | 6.85×10^{-6} | 0.486 |
| Mother's age (quadratic term) | 8.28×10^{-10} | 1.68×10^{-9} | 0.493 |
| Child's birth order (linear term) | 1.60×10^{-4} | 3.87×10^{-4} | 0.413 |
| Child's birth order (quadratic term) | 8.55×10^{-6} | 2.24×10^{-5} | 0.382 |
| Wealth index | 1.74×10^{-6} | 4.05×10^{-6} | 0.430 |
| 0 -rural; 1 - urban | 1.27×10^{-5} | 4.21×10^{-5} | 0.303 |
| Mother's education level | 4.56×10^{-6} | 1.20×10^{-5} | 0.380 |
| Child is boy | 7.12×10^{-6} | 1.91×10^{-5} | 0.373 |
| Mother is married or living together | 1.83×10^{-5} | 4.96×10^{-5} | 0.370 |
| Antenatal care indicator | 9.63×10^{-6} | 2.16×10^{-5} | 0.447 |

Note that in our multiple imputation procedure, the variance ratios are all less than 0.5, indicating that for each regressor the variance due to missing data (between-imputation variance) is much less than the average estimated squared standard error over the 500 imputed data sets. More replications of imputation (larger m) will more sufficiently reduce the variation due to missingness and therefore lead to more reliable estimation (Rubin, 1987; Schafer, 1999). We take a sufficiently large

number of replications $m = 500$ to ensure that the variance due to missingness has been sufficiently controlled.

4.5.7 Design of the sensitivity analyses

In Section "Sensitivity analyses" of the main text, we very briefly described three perspectives on how potential unobserved covariates that cannot be adjusted by matching may bias the estimated effect in a difference-in-differences study. Here we give more detailed descriptions of them with connections to our study for reference:

- Perspective 1: The potential violation of the unconfoundedness assumption (Rosenbaum and Rubin, 1983b; Heckman and Robb, 1985). Roughly speaking, the unconfoundedness assumption states that, after adjusting for observed covariates (measured confounders), there are no differential trends over time of any characteristics, other than the intervention itself, between the treated group and the control group, that may be correlated with their outcomes. This assumption may be violated if there is selection bias on unobserved covariates across time (Heckman and Robb, 1985; Heckman et al., 1997) such that there are differences in these observed covariates of the treated group and the control group which impact their trends in the outcome (Ashenfelter and Card, 1984; Doyle et al., 2018). For example, in our study, the unconfoundedness assumption can be violated if the sharp drops in malaria prevalence experienced by some areas could be explained by the changes of some unobserved characteristics over time that could also predict the low birth weight rate.
- Perspective 2: The potential violation of the parallel trend assumption in a

difference-in-differences study (Card and Krueger, 2000; Angrist and Pischke, 2008; Hasegawa et al., 2019; Basu and Small, 2020). Recall that the parallel trend assumption behind a difference-in-differences study states that, in the absence of the treatment (i.e., intervention), after adjusting for relevant covariates, the outcome trajectory of the treated group would follow a parallel trend with that of the control group. Therefore, to make the parallel trend assumption more likely to hold, ideally each observed or unobserved covariate should be well balanced (i.e., follow a common trajectory) between the treated group and the control group, before and after the intervention. Matching can balance observed covariates by ensuring each covariate follows a common trajectory in the treated and control groups. However, matching cannot directly adjust for unobserved covariates and their trajectories among the treated and control groups may differ and correspondingly the parallel trend assumption may not hold.

- Perspective 3: A difference-in-differences study may be biased if there is an event that is more (or less) likely to occur as the treatment (i.e., intervention) happens in the treated group, but, unlike the case discussed in Section “Motivation and overview of our approach” of the main text, the occurrence probability of this event cannot be fully captured by observed covariates. In this case, if this event can affect the outcome, its contribution to the outcome will be more (or less) substantial within the treated group after the treatment (i.e., intervention) than that within the control group (Shadish, 2010; West and Thoemmes, 2010). For example, areas experiencing sharp drops in malaria prevalence might also be more likely to experience other events (e.g., sharp drops in the prevalence of other infectious diseases) that can contribute to

decreasing the low birth weight rate.

We use an omitted variable sensitivity analysis approach (Rosenbaum and Rubin, 1983a; Imbens, 2003; Ichino et al., 2008; Zhang and Small, 2020) to evaluate the sensitivity of the results of our primary analysis to potential hidden bias caused by unobserved covariates. Specifically, we propose the following sensitivity analysis model (4.3) which extends Model (4.1) by considering a hypothetical unobserved covariate (unmeasured confounding variable or event) U that is correlated with both the low malaria prevalence indicator and the low birth weight indicator. Let U_{ij} denote the exact value of U for individual j in cluster i , we consider:

$$\begin{aligned} \mathbb{P}(Y_{ij} = 1 \mid i, \mathbf{X}_{ij}, U_{ij}) &= k_0 + k_1 \cdot \mathbb{1}(i \text{ is a low malaria prevalence cluster}) \\ &\quad + k_2 \cdot \mathbb{1}(i \text{ is a late year cluster}) \\ &\quad + k_3 \cdot \mathbb{1}(i \text{ is from a high-low pair of clusters}) \\ &\quad + \beta^T \mathbf{X}_{ij} + \lambda \cdot U_{ij}, \end{aligned} \quad (4.3)$$

with two error terms $\alpha_i \sim \mathcal{N}(0, \sigma_0)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_1)$, and U_{ij} follows a Bernoulli distribution (taking value 0 or 1) with

$$\begin{aligned} \mathbb{P}(U_{ij} = 1) &= 50\% + p_1\% \cdot \mathbb{1}(i \text{ is a low malaria prevalence cluster}) \\ &\quad + p_2\% \cdot \mathbb{1}(\text{the observed or the imputed } Y_{ij} = 1). \end{aligned} \quad (4.4)$$

In Model (4.3) along with the corresponding data generating model (4.4) of the unobserved covariate U , the (p_1, p_2) are sensitivity parameters of which the unit is a percentage point. Prespecifying a positive (or negative) p_1 corresponds to a positive (or negative) correlation between the unobserved covariate U and the low malaria prevalence indicator, and prespecifying a positive (or negative) p_2 corre-

sponds to a positive (or negative) correlation between the unobserved covariate U and the low birth weight indicator. It is clear that a larger magnitude of p_1 (or p_2) corresponds to a larger magnitude of correlation between U and the low malaria prevalence indicator (or the low birth weight indicator). We now discuss how the proposed sensitivity analysis model helps to address concerns about the potential hidden bias from Perspectives 1–3 listed above:

- For Perspective 1: The proposed sensitivity analysis model covers Perspective 1 by considering a hypothetical unobserved covariate U such that it is correlated with both the low malaria prevalence indicator (i.e., the indicator for units who have experienced sharp drops in malaria prevalence) (by prespecifying various p_1) and the low birth weight indicator (by prespecifying various p_2). With the unobserved covariate U , the unconfoundedness assumption may be violated as matching can only adjust for observed covariates but cannot directly adjust for unobserved covariates.
- For Perspective 2: The proposed sensitivity analysis model also covers Perspective 2 by including the unobserved covariate U in the final outcome model. This is because by setting a non-zero p_1 , the distributions of U between high-low and high-high pairs of clusters will be imbalanced (i.e., will not follow a common trajectory). Meanwhile, by setting a non-zero p_2 (corresponds to a non-zero λ in Model (4.3)), the imbalances of U across the treated and controls will make the outcome trend of the high-low pairs of clusters (i.e., the treated group) in the absence of the treatment deviate from a parallel trend with that of the high-high pairs (i.e., the control group).
- For Perspective 3: When setting $p_1 \neq 0$, the hypothetical unobserved covariate U in our sensitivity analysis model can also be regarded as some event

of which the occurrence probability varies across the treated group and the control group and is not directly associated with observed covariates. Meanwhile, by setting some $p_2 \neq 0$, the contribution of that event to the low birth weight rate differs across the treated group and the control group as that event occurs more (or less) frequently in the treated group. Therefore, our sensitivity analyses also cover Perspective 3 of the potential hidden bias.

After setting up the sensitivity analysis model (4.3), the detailed sensitivity analysis procedure is as follows. For each pair of prespecified sensitivity parameters (p_1, p_2) and for each imputed data set (500 in total) obtained from Step 3 (the multiple imputation stage), we generate the value of U_{ij} for each individual j in cluster i according to Model (4.4) and calculate the corresponding point estimate and estimated standard error of the coefficient of the low malaria prevalence indicator under Model (4.3). Similarly to the primary analysis, for each pair of prespecified (p_1, p_2) , the corresponding estimated causal effect of reduced malaria burden on the low birth weight rate is the mean value of the 500 estimated coefficients on the low malaria prevalence indicator obtained from 500 runs of Model (4.3). The corresponding p-value and 95% CIs can also be obtained via applying Rubin's rules with treating the imputed U as an usual regressor in Model (4.3). We conduct the above procedure for various (p_1, p_2) and examine how the results differ from those in the primary analysis.

4.5.8 Detailed results of the sensitivity analyses

When reporting the sensitivity analyses for the coefficient of the low malaria prevalence indicator under the sensitivity analysis model (4.3) with various prespecified values of the sensitivity parameters (p_1, p_2) , we divide the results into the follow-

ing four cases:

- Case 1: $p_1 > 0, p_2 > 0$. That is, the hypothetical unobserved covariate U is positively correlated with both the low malaria prevalence indicator (i.e., the indicator for units who have experienced sharp drops in malaria prevalence) and the low birth weight indicator (i.e., the outcome variable).
- Case 2: $p_1 > 0, p_2 < 0$. That is, the hypothetical unobserved covariate U is positively correlated with the low malaria prevalence indicator while it is negatively correlated with the low birth weight indicator.
- Case 3: $p_1 < 0, p_2 > 0$. That is, the hypothetical unobserved covariate U is negatively correlated with the low malaria prevalence indicator while it is positively correlated with the low birth weight indicator.
- Case 4: $p_1 < 0, p_2 < 0$. That is, the hypothetical unobserved covariate U is negatively correlated with both the low malaria prevalence indicator and the low birth weight indicator.

We report the results of the sensitivity analyses in Table 4.11 of Appendix 4.5.7. Specifically, for each (p_1, p_2) , we report the point estimate, the 95% CI, and the p -value (under null effect) of the low malaria prevalence indicator under Model (4.3) in which the hypothetical unobserved covariate U_{ij} is generated from Model (4.4) within each imputed data set.

We list the interpretations of the results in Table 4.11 of Appendix 4.5.7 case by case:

- Cases 1 and 4: In these two cases, the magnitude of the estimated treatment effect obtained from the primary analysis assuming no observed co-

variates (1.48 percentage points reduction, listed in Table 4.6 of the main text) is smaller than that obtained from the sensitivity analyses in which the unobserved covariate U is taken into account. This implies that if the unobserved covariate more (or less) frequently appears in the treated group and predicts the outcome in the opposite (or same) direction as the treatment does, the primary analysis tends to underestimate the actual treatment effect. This pattern agrees with the previous literature on sensitivity analyses (Gastwirth et al., 1998; Rosenbaum and Silber, 2009). However, as shown in Table 4.11 of Appendix 4.5.7, the magnitude of this potential estimation bias is estimated to be no greater than $|-1.83 - (-1.48)| = 0.35$ percentage points as long as $p_1, p_2 \in (0, 10]$ percentage points or $p_1, p_2 \in [-10, 0)$ percentage points.

- Cases 2 and 3: In these two cases, the magnitude of the estimated treatment effect obtained from the primary analysis is smaller than that obtained from the sensitivity analyses with U taken into account. This implies that if the unobserved covariate more (or less) frequently appears in the treated group and predicts the outcome in the same (or opposite) direction as the treatment does, the primary analysis tends to overestimate the actual treatment effect. This pattern also agrees with the previous literature on sensitivity analyses (Gastwirth et al., 1998; Rosenbaum and Silber, 2009). However, as shown in Table 4.11 of Appendix 4.5.7, the magnitude of this potential estimation bias is estimated to be no greater than $|-1.13 - (-1.48)| = 0.35$ percentage points as long as $|p_1| \leq 10$ percentage points and $|p_2| \leq 10$ percentage points.

Table 4.11: The results of the sensitivity analyses for the coefficient of the low malaria prevalence indicator under various sensitivity parameters (p_1, p_2) divided into the four cases: Case 1: $p_1 > 0, p_2 > 0$; Case 2: $p_1 > 0, p_2 < 0$; Case 3: $p_1 < 0, p_2 > 0$; Case 4: $p_1 < 0, p_2 < 0$. The unit of estimates and CIs is a percentage point.

| Case 1 | $p_2 = 5.0$ | | | $p_2 = 10.0$ | | |
|---------------|--------------|---------------|---------|---------------|---------------|---------|
| | Estimate | 95% CI | p-value | Estimate | 95% CI | p-value |
| $p_1 = 2.5$ | -1.52 | [-3.74, 0.70] | 0.179 | -1.56 | [-3.77, 0.66] | 0.168 |
| $p_1 = 5.0$ | -1.57 | [-3.79, 0.66] | 0.167 | -1.65 | [-3.86, 0.57] | 0.145 |
| $p_1 = 7.5$ | -1.61 | [-3.83, 0.61] | 0.156 | -1.73 | [-3.95, 0.48] | 0.125 |
| $p_1 = 10.0$ | -1.65 | [-3.88, 0.57] | 0.145 | -1.82 | [-4.04, 0.40] | 0.107 |
| Case 2 | $p_2 = -5.0$ | | | $p_2 = -10.0$ | | |
| | Estimate | 95% CI | p-value | Estimate | 95% CI | p-value |
| $p_1 = 2.5$ | -1.44 | [-3.66, 0.78] | 0.204 | -1.39 | [-3.62, 0.83] | 0.219 |
| $p_1 = 5.0$ | -1.40 | [-3.62, 0.83] | 0.218 | -1.31 | [-3.53, 0.92] | 0.249 |
| $p_1 = 7.5$ | -1.35 | [-3.58, 0.87] | 0.234 | -1.22 | [-3.44, 1.00] | 0.282 |
| $p_1 = 10.0$ | -1.31 | [-3.53, 0.92] | 0.250 | -1.13 | [-3.36, 1.09] | 0.318 |
| Case 3 | $p_2 = 5.0$ | | | $p_2 = 10.0$ | | |
| | Estimate | 95% CI | p-value | Estimate | 95% CI | p-value |
| $p_1 = -2.5$ | -1.44 | [-3.66, 0.78] | 0.204 | -1.39 | [-3.61, 0.83] | 0.219 |
| $p_1 = -5.0$ | -1.39 | [-3.61, 0.83] | 0.219 | -1.30 | [-3.52, 0.91] | 0.249 |
| $p_1 = -7.5$ | -1.35 | [-3.57, 0.87] | 0.234 | -1.22 | [-3.43, 1.00] | 0.282 |
| $p_1 = -10.0$ | -1.31 | [-3.53, 0.92] | 0.250 | -1.13 | [-3.35, 1.09] | 0.319 |
| Case 4 | $p_2 = -5.0$ | | | $p_2 = -10.0$ | | |
| | Estimate | 95% CI | p-value | Estimate | 95% CI | p-value |
| $p_1 = -2.5$ | -1.52 | [-3.75, 0.70] | 0.179 | -1.56 | [-3.79, 0.66] | 0.168 |
| $p_1 = -5.0$ | -1.57 | [-3.79, 0.66] | 0.167 | -1.65 | [-3.87, 0.57] | 0.146 |
| $p_1 = -7.5$ | -1.61 | [-3.84, 0.61] | 0.156 | -1.74 | [-3.96, 0.49] | 0.126 |
| $p_1 = -10.0$ | -1.66 | [-3.88, 0.57] | 0.145 | -1.83 | [-4.05, 0.40] | 0.108 |

Bibliography

- Abadie, A. (2005). Semiparametric difference-in-differences estimators. *The Review of Economic Studies*, 72(1):1–19.
- Adamu, A. L., Crampin, A., Kayuni, N., Amberbir, A., Koole, O., Phiri, A., Nyirenda, M., and Fine, P. (2017). Prevalence and risk factors for anemia severity and type in malawian men and women: urban and rural differences. *Population Health Metrics*, 15(1):1–15.
- Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Arnold, B. F. and Ercumen, A. (2016). Negative control outcomes: a tool to detect bias in randomized trials. *JAMA*, 316(24):2597–2598.
- Ashenfelter, O. C. and Card, D. (1984). Using the longitudinal structure of earnings to estimate the effect of training programs.
- Athey, S. and Imbens, G. W. (2006). Identification and inference in nonlinear difference-in-differences models. *Econometrica*, 74(2):431–497.
- Athey, S. and Imbens, G. W. (2017). The econometrics of randomized experiments. In *Handbook of Economic Field Experiments*, volume 1, pages 73–140. Elsevier.
- Avchen, R. N., Scott, K. G., and Mason, C. A. (2001). Birth weight and school-age disabilities: a population-based study. *American Journal of Epidemiology*, 154(10):895–901.
- Ayele, D. G., Zewotir, T. T., and Mwambi, H. G. (2013). The risk factor indicators of malaria in ethiopia. *International Journal of Medicine and Medical Sciences*, 5(7):335–347.
- Baragatti, M., Fournet, F., Henry, M.-C., Assi, S., Ouedraogo, H., Rogier, C., and Salem, G. (2009). Social and environmental malaria risk factors in urban areas of ouagadougou, burkina faso. *Malaria Journal*, 8(1):1–14.
- Basu, P. and Small, D. S. (2020). Constructing a more closely matched control group in a difference-in-differences analysis: its effect on history interacting with group bias. *Observational Studies*, 6:103–130.
- Beesley, L. J. and Mukherjee, B. (2020). Statistical inference for association stud-

- ies using electronic health records: handling both selection bias and outcome misclassification. *Biometrics*, 78(1):214–226.
- Bhatt, S., Weiss, D., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K., Moyes, C., Henry, A., Eckhoff, P., Wenger, E., Briët, O., Penny, M., Smith, T., Bennett, A., Yukich, J., Eisele, T., Griffin, J., Fergus, C., Lynch, M., Lindgren, F., Cohen, J., Murray, C., Smith, D., Hay, S., Cibulskis, R., and Gething, P. (2015). The effect of malaria control on plasmodium falciparum in africa between 2000 and 2015. *Nature*, 526(7572):207.
- Billingsley, P. (1995). *Measure and Probability*. A Wiley-Interscience Publication, John Wiley & Sons.
- Blanc, A. K. and Wardlaw, T. (2005). Monitoring low birth weight: an evaluation of international estimates and an updated estimation procedure. *Bulletin of the World Health Organization*, 83:178–185d.
- Bloss, E., Wainaina, F., and Bailey, R. C. (2004). Prevalence and predictors of underweight, stunting, and wasting among children aged 5 and under in western kenya. *Journal of Tropical Pediatrics*, 50(5):260–270.
- Boerma, J. T., Weinstein, K., Rutstein, S. O., and Sommerfelt, A. E. (1996). Data on birth weight in developing countries: can surveys help? *Bulletin of the World Health Organization*, 74(2):209.
- Boyle, E. H., King, M., and Sobek, M. (2019). *Minnesota Population Center and ICF International*.
- Brooke, O. G., Anderson, H. R., Bland, J. M., Peacock, J. L., and Stewart, C. M. (1989). Effects on birth weight of smoking, alcohol, caffeine, socioeconomic factors, and psychosocial stress. *BMJ*, 298(6676):795–801.
- Brown, K. H., Black, R. E., Becker, S., et al. (1982). Seasonal changes in nutritional status and the prevalence of malnutrition in a longitudinal study of young children in rural bangladesh. *Am J Clin Nutr*, 36(2):303–13.
- Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. Chapman and Hall/CRC.
- Burer, S. and Saxena, A. (2012). The MILP road to MIQCP. In *Mixed Integer Non-linear Programming*, pages 373–405. Springer.
- Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algo-

- rithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208.
- Card, D. and Krueger, A. B. (2000). Minimum wages and employment: a case study of the fast-food industry in new jersey and pennsylvania: reply. *American Economic Review*, 90(5):1397–1420.
- Carpenter, J. and Kenward, M. (2012). *Multiple imputation and its application*. John Wiley & Sons.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC.
- Catania, J. A., McDermott, L. J., and Pollack, L. M. (1986). Questionnaire response bias and face-to-face interview sample bias in sexuality research.
- Charalambous, C. and Conn, A. (1978). An efficient method to solve the minimax problem directly. *SIAM Journal on Numerical Analysis*, 15(1):162–187.
- Chen, X.-K., Wen, S. W., Sun, L.-M., Yang, Q., Walker, M. C., and Krewski, D. (2009). Recent oral contraceptive use and adverse birth outcomes. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 144(1):40–43.
- Conforti, M., Cornuéjols, G., and Zambelli, G. (2014). *Integer Programming*, volume 271. Springer.
- Corsi, D. J., Neuman, M., Finlay, J. E., and Subramanian, S. (2012). Demographic and health surveys: a profile. *International Journal of Epidemiology*, 41(6):1602–1613.
- Cox, D. R. and Snell, E. J. (2018). *Analysis of Binary Data*. Routledge.
- Darteh, E. K. M., Acquah, E., and Kumi-Kyereme, A. (2014). Correlates of stunting among children in ghana. *BMC Public Health*, 14(1):1–7.
- de Bernabé, J. V., Soriano, T., Albaladejo, R., Juarranz, M., Calle, M. E., Martínez, D., and Domínguez-Rojas, V. (2004). Risk factors for low birth weight: a review. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, 116(1):3–15.
- Dellicour, S., Tatem, A. J., Guerra, C. A., Snow, R. W., and ter Kuile, F. O. (2010). Quantifying the number of pregnancies at risk of malaria in 2007: a demographic study. *PLoS Medicine*, 7(1):e1000221.
- DHS (2019). Methodology - collecting geographic data. *Demographic and Health Surveys (DHS)*.

- Dimick, J. B. and Ryan, A. M. (2014). Methods for evaluating changes in health care policy: the difference-in-differences approach. *JAMA*, 312(22):2401–2402.
- Doyle, O., Hegarty, M., and Owens, C. (2018). Population-based system of parenting support to reduce the prevalence of child social, emotional, and behavioural problems: difference-in-differences study. *Prevention Science*, 19(6):772–781.
- Eisele, T. P., Larsen, D. A., Anglewicz, P. A., Keating, J., Yukich, J., Bennett, A., Hutchinson, P., and Steketee, R. W. (2012). Malaria prevention in pregnancy, birthweight, and neonatal mortality: a meta-analysis of 32 national cross-sectional datasets in africa. *The Lancet Infectious Diseases*, 12(12):942–949.
- Ertefaie, A., Small, D. S., and Rosenbaum, P. R. (2018). Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. *Journal of the American Statistical Association*, 113(523):1122–1134.
- Fakhouri, T. H., Martin, C. B., Chen, T.-C., Akinbami, L. J., Ogden, C. L., Paulose-Ram, R., Riddles, M. K., Van de Kerckhove, W., Roth, S. B., Clark, J., Mohadjer, L. K., and Fay, R. E. (2020). An investigation of nonresponse bias and survey location variability in the 2017-2018 national health and nutrition examination survey. *Vital and Health statistics. Series 2, Data Evaluation and Methods Research*, (185):1–36.
- Fink, G., Günther, I., and Hill, K. (2011). The effect of water and sanitation on child health: evidence from the demographic and health surveys 1986–2007. *International Journal of Epidemiology*, 40(5):1196–1204.
- Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh and London: Oliver and Boyd.
- Fogarty, C. B. (2018). On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1035–1056.
- Fogarty, C. B., Lee, K., Kelz, R. R., and Keele, L. J. (2021). Biased encouragements and heterogeneous effects in an instrumental variable study of emergency general surgical outcomes. *Journal of the American Statistical Association*, 116(536):1625–1636.
- Fogarty, C. B., Mikkelsen, M. E., Gaieski, D. F., and Small, D. S. (2016). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, 111(514):447–458.

- Fogarty, C. B., Shi, P., Mikkelsen, M. E., and Small, D. S. (2017). Randomization inference and sensitivity analysis for composite null hypotheses with binary outcomes in matched observational studies. *Journal of the American Statistical Association*, 112(517):321–331.
- 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.
- Fowkes, F. J., Davidson, E., Moore, K. A., McGready, R., and Simpson, J. A. (2020). The invisible burden of malaria-attributable stillbirths. *The Lancet*, 395(10220):268.
- Fraser, A. M., Brockert, J. E., and Ward, R. H. (1995). Association of young maternal age with adverse reproductive outcomes. *New England Journal of Medicine*, 332(17):1113–1118.
- Gałecki, A. and Burzykowski, T. (2013). Linear mixed-effects model. In *Linear Mixed-Effects Models Using R*, pages 245–273. Springer.
- Garrett, J. L. and Ruel, M. T. (2005). Stunted child–overweight mother pairs: prevalence and association with economic development and urbanization. *Food and nutrition bulletin*, 26(2):209–221.
- Gastwirth, J. L., Krieger, A. M., and Rosenbaum, P. R. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika*, 85(4):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.
- Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383.
- Gemperli, A., Vounatsou, P., Kleinschmidt, I., Bagayoko, M., Lengeler, C., and Smith, T. (2004). Spatial patterns of infant mortality in mali: the effect of malaria endemicity. *American Journal of Epidemiology*, 159(1):64–72.
- Gething, P., Hay, S., and Weiss, D. (2020). The invisible burden of malaria-attributable stillbirths—authors’ reply. *The Lancet*, 395(10220):268–269.
- Gilbert, R., Martin, R. M., Donovan, J., Lane, J. A., Hamdy, F., Neal, D. E., and Met-

- calfe, C. (2016). Misclassification of outcome in case–control studies: methods for sensitivity analysis. *Statistical Methods in Medical Research*, 25(5):2377–2393.
- Goetgeluk, S. and Vansteelandt, S. (2008). Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics*, 64(3):772–780.
- Grace, K., Davenport, F., Hanson, H., Funk, C., and Shukla, S. (2015). Linking climate change and health outcomes: Examining the relationship between temperature, precipitation and birth weight in africa. *Global Environmental Change*, 35:125–137.
- Gurobi Optimization, LLC (2022). Gurobi Optimizer Reference Manual.
- Guyatt, H. L. and Snow, R. W. (2004). Impact of malaria during pregnancy on low birth weight in sub-saharan africa. *Clinical Microbiology Reviews*, 17(4):760–769.
- Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:361–374.
- Hansen, B. B. (2004). Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association*, 99(467):609–618.
- 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, B. B., Rosenbaum, P. R., and Small, D. S. (2014). Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *Journal of the American Statistical Association*, 109(505):133–144.
- Harris, N. S., Crawford, P. B., Yangzom, Y., Pinzo, L., Gyaltzen, P., and Hudes, M. (2001). Nutritional and health status of tibetan children living at high altitudes. *New England Journal of Medicine*, 344(5):341–347.
- Hasegawa, R. B., Webster, D. W., and Small, D. S. (2019). Evaluating missouri’s handgun purchaser law: a bracketing method for addressing concerns about history interacting with group. *Epidemiology*, 30(3):371–379.
- Hay, S. I. and Snow, R. W. (2006). The malaria atlas project: developing global maps of malaria risk. *PLoS Medicine*, 3(12):e473.
- Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

- Heckman, J. J. and Robb, R. J. (1985). Alternative methods for evaluating the impact of interventions: An overview. *Journal of Econometrics*, 30(1-2):239–267.
- Heitjan, D. F. and Basu, S. (1996). Distinguishing “missing at random” and “missing completely at random”. *The American Statistician*, 50(3):207–213.
- Heller, R., Rosenbaum, P. R., and Small, D. S. (2009). Split samples and design sensitivity in observational studies. *Journal of the American Statistical Association*, 104(487):1090–1101.
- Heng, S., Small, D. S., and Rosenbaum, P. R. (2020). Finding the strength in a weak instrument in a study of cognitive outcomes produced by catholic high schools. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(3):935–958.
- Hernán, M. A. and Robins, J. M. (2020). Causal Inference: What If.
- Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236.
- 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. *The Annals of Applied Statistics*, 4(2):849–870.
- Howard, S. R. and Pimentel, S. D. (2019). The uniform general signed rank test and its design sensitivity. *arXiv preprint arXiv:1904.08895*.
- Huynh, B.-T., Cottrell, G., Cot, M., and Briand, V. (2015). Burden of malaria in early pregnancy: a neglected problem? *Clinical Infectious Diseases*, 60(4):598–604.
- Huynh, B.-T., Fievet, N., Gbaguidi, G., Dechavanne, S., Borgella, S., Guézo-Mévo, B., Massougboji, A., Ndam, N. T., Deloron, P., and Cot, M. (2011). Influence of the timing of malaria infection during pregnancy on birth weight and on maternal anemia in benin. *The American Journal of Tropical Medicine and Hygiene*, 85(2):214–220.
- ICF (2019). 2004-2017. demographic and health surveys (various) [datasets]. funded by usaid. rockville, maryland: Icf [distributor]. Technical report.
- Ichino, A., Mealli, F., and Nannicini, T. (2008). From temporary help jobs to permanent employment: what can we learn from matching estimators and their sensitivity? *Journal of Applied Econometrics*, 23(3):305–327.
- 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 Rubin, D. B. (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Kayentao, K., Garner, P., van Eijk, A. M., Naidoo, I., Roper, C., Mulokozi, A., MacArthur, J. R., Luntamo, M., Ashorn, P., Doumbo, O. K., and ter Kuile, F. O. (2013). Intermittent preventive therapy for malaria during pregnancy using 2 vs 3 or more doses of sulfadoxine-pyrimethamine and risk of low birth weight in africa: systematic review and meta-analysis. *JAMA*, 309(6):594–604.
- Keele, L. and Quinn, K. M. (2017). Bayesian sensitivity analysis for causal effects from 2×2 tables in the presence of unmeasured confounding with application to presidential campaign visits. *The Annals of Applied Statistics*, 11(4):1974–1997.
- Kennedy, E., Gray, N., Azzopardi, P., and Creati, M. (2011). Adolescent fertility and family planning in east asia and the pacific: a review of dhs reports. *Reproductive Health*, 8(1):11.
- Knäuper, B. and Wittchen, H.-U. (1994). Diagnosing major depression in the elderly: evidence for response bias in standardized diagnostic interviews? *Journal of Psychiatric Research*, 28(2):147–164.
- Kramer, M. S. (1987). Determinants of low birth weight: methodological assessment and meta-analysis. *Bulletin of the World Health Organization*, 65(5):663.
- Krefis, A. C., Schwarz, N. G., Nkrumah, B., Acquah, S., Loag, W., Sarpong, N., Adu-Sarkodie, Y., Ranft, U., and May, J. (2010). Principal component analysis of socioeconomic factors and their association with malaria in children from the ashanti region, ghana. *Malaria Journal*, 9(1):1–7.
- Küchenhoff, H., Mwalili, S. M., and Lesaffre, E. (2006). A general method for dealing with misclassification in regression: the misclassification simex. *Biometrics*, 62(1):85–96.
- Larsen, D. A., Grisham, T., Slawsky, E., and Narine, L. (2017). An individual-level meta-analysis assessing the impact of community-level sanitation access on child stunting, anemia, and diarrhea: Evidence from dhs and mics surveys. *PLoS Neglected Tropical Diseases*, 11(6):e0005591.
- Lee, J. and Leyffer, S. (2011). *Mixed Integer Nonlinear Programming*, volume 154. Springer Science & Business Media.

- Li, X. and Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769.
- Lubovsky, O., Liebergall, M., Mattan, Y., Weil, Y., and Mosheiff, R. (2005). Early diagnosis of occult hip fractures: Mri versus ct scan. *Injury*, 36(6):788–792.
- Lucia, M. S., Epstein, J. I., Goodman, P. J., Darke, A. K., Reuter, V. E., Civantos, F., Tangen, C. M., Parnes, H. L., Lippman, S. M., La Rosa, F. G., et al. (2007). Finasteride and high-grade prostate cancer in the prostate cancer prevention trial. *Journal of the National Cancer Institute*, 99(18):1375–1383.
- Lyles, R. H., Williamson, J. M., Lin, H.-M., and Heilig, C. M. (2005). Extending mcnemar’s test: Estimation and inference when paired binary outcome data are misclassified. *Biometrics*, 61(1):287–294.
- Magder, L. S. and Hughes, J. P. (1997). Logistic regression when the outcome is measured with uncertainty. *American Journal of Epidemiology*, 146(2):195–203.
- Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. 22(4):719–748.
- MAP (2020). Malaria atlas project. *The MAP Group*.
- Margot, F. (2010). Symmetry in integer linear programming. In *50 Years of Integer Programming 1958-2008*, pages 647–686. Springer.
- Mayor, A., Bardají, A., Macete, E., Nhampossa, T., Fonseca, A. M., González, R., Maculuve, S., Cisteró, P., Rupérez, M., Campo, J., Vala, A., Sigaúque, B., Jiménez, A., Machevo, S., de la Fuente, L., Nhama, A., Luis, L., Aponte, J. J., Acácio, S., Nhacolo, A., Chitnis, C., Dobaño, C., Sevene, E., Alonso, P. L., and Menéndez, C. (2015). Changing trends in p. falciparum burden, immunity, and disease in pregnancy. *New England Journal of Medicine*, 373(17):1607–1617.
- 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.
- McCormick, M. C., Brooks-Gunn, J., Workman-Daniels, K., Turner, J., and Peckham, G. J. (1992). The health and developmental status of very low—birth-weight children at school age. *JAMA*, 267(16):2204–2208.
- McGregor, I. A., Wilson, M., and Billewicz, W. (1983). Malaria infection of the placenta in the gambia, west africa; its incidence and relationship to stillbirth, birth-

- weight and placental weight. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 77(2):232–244.
- Menendez, C., Ordi, J., Ismail, M., Ventura, P., Aponte, J., Kahigwa, E., Font, F., and Alonso, P. (2000). The impact of placental malaria on gestational age and birth weight. *The Journal of Infectious Diseases*, 181(5):1740–1745.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, pages 538–558.
- Metselaar, D. and Van Thiel, P. (1959). Classification of malaria. *Tropical and Geographical Medicine*, 11(2):157–61.
- Mitra, N. and Heitjan, D. F. (2007). Sensitivity of the hazard ratio to nonignorable treatment assignment in an observational study. *Statistics in Medicine*, 26(6):1398–1414.
- Neuman, M. D., Rosenbaum, P. R., Ludwig, J. M., Zubizarreta, J. R., and Silber, J. H. (2014). Anesthesia technique, mortality, and length of stay after hip fracture surgery. *JAMA*, 311(24):2508–2517.
- Neyman, J. S. (1923). On the application of probability theory to agricultural experiments. essay on principles. section 9. (translated and edited by D. M. Dabrowska and T. P. Speed). *Statistical Science*, (1990) 5:465–480.
- NIH (2013). *Prostate Cancer Prevention Trial (PCPT): Questions and Answers*. National Cancer Institute at the National Institutes of Health.
- Null, C., Stewart, C. P., Pickering, A. J., Dentz, H. N., Arnold, B. F., Arnold, C. D., Benjamin-Chung, J., Clasen, T., Dewey, K. G., Fernald, L. C., et al. (2018). Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural kenya: a cluster-randomised controlled trial. *The Lancet Global Health*, 6(3):e316–e329.
- Ogedegbe, G., Pickering, T. G., Clemow, L., Chaplin, W., Spruill, T. M., Albanese, G. M., Eguchi, K., Burg, M., and Gerin, W. (2008). The misdiagnosis of hypertension: the role of patient anxiety. *Archives of Internal Medicine*, 168(22):2459–2465.
- Padhi, B. K., Baker, K. K., Dutta, A., Cumming, O., Freeman, M. C., Satpathy, R., Das, B. S., and Panigrahi, P. (2015). Risk of adverse pregnancy outcomes among women practicing poor sanitation in rural india: a population-based prospective cohort study. *PLoS Medicine*, 12(7):e1001851.

- Paneth, N. S. (1995). The problem of low birth weight. *The Future of Children*, pages 19–34.
- Phuka, J. C., Maleta, K., Thakwalakwa, C., Cheung, Y. B., Briend, A., Manary, M. J., and Ashorn, P. (2008). Complementary feeding with fortified spread and incidence of severe stunting in 6-to 18-month-old rural malawians. *Archives of Pediatrics & Adolescent Medicine*, 162(7):619–626.
- Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association*, 110(510):515–527.
- Quade, D., Lachenbruch, P. A., Whaley, F. S., McCLISH, D. K., and Haley, R. W. (1980). Effects of misclassifications on statistical inferences in epidemiology. *American Journal of Epidemiology*, 111(5):503–515.
- R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Radeva-Petrova, D., Kayentao, K., ter Kuile, F. O., Sinclair, D., and Garner, P. (2014). Drugs for preventing malaria in pregnant women in endemic areas: any drug regimen versus placebo or no treatment. *Cochrane Database of Systematic Reviews*, (10).
- Redman, M. W., Tangen, C. M., Goodman, P. J., Lucia, M. S., Coltman, C. A., and Thompson, I. M. (2008). Finasteride does not increase the risk of high-grade prostate cancer: a bias-adjusted modeling approach. *Cancer Prevention Research*, 1(3):174–181.
- Richards, M., Hardy, R., Kuh, D., and Wadsworth, M. E. (2001). Birth weight and cognitive function in the british 1946 birth cohort: longitudinal population based study. *BMJ*, 322(7280):199–203.
- Roberts, D. and Matthews, G. (2016). Risk factors of malaria in children under the age of five years old in uganda. *Malaria Journal*, 15(1):1–11.
- Robles, A. and Goldman, N. (1999). Can accurate data on birthweight be obtained from health interview surveys? *International Journal of Epidemiology*, 28(5):925–931.
- Rogerson, S. J., Mwapasa, V., and Meshnick, S. R. (2007). Malaria in pregnancy: linking immunity and pathogenesis to prevention. In *Defining and Defeating the Intolerable Burden of Malaria III: Progress and Perspectives: Supplement to Volume*

- 77 (6) of *American Journal of Tropical Medicine and Hygiene*. American Society of Tropical Medicine and Hygiene.
- Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society: Series A (General)*, 147(5):656–666.
- Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, 84(408):1024–1032.
- Rosenbaum, P. R. (1991). A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(3):597–610.
- Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327.
- Rosenbaum, P. R. (2002b). *Observational Studies*. Springer.
- Rosenbaum, P. R. (2004). Design sensitivity in observational studies. *Biometrika*, 91(1):153–164.
- Rosenbaum, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics*, 63(2):456–464.
- Rosenbaum, P. R. (2010). *Design of Observational Studies*, volume 10. Springer.
- Rosenbaum, P. R. (2011). A new u-statistic with superior design sensitivity in matched observational studies. *Biometrics*, 67(3):1017–1027.
- Rosenbaum, P. R. (2012). Testing one hypothesis twice in observational studies. *Biometrika*, 99(4):763–774.
- Rosenbaum, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics*, 69(1):118–127.
- Rosenbaum, P. R. (2014). Weighted m-statistics with superior design sensitivity in matched observational studies with multiple controls. *Journal of the American Statistical Association*, 109(507):1145–1158.
- Rosenbaum, P. R. (2017). *Observation and Experiment: An Introduction to Causal Inference*. Harvard University Press.
- Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved

- binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218.
- Rosenbaum, P. R. and Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.
- Rosenbaum, P. R. and Silber, J. H. (2008). Aberrant effects of treatment. *Journal of the American Statistical Association*, 103(481):240–247.
- Rosenbaum, P. R. and Silber, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *Journal of the American Statistical Association*, 104(488):1398–1405.
- Rosenbaum, P. R. and Small, D. S. (2017). An adaptive mantel–haenszel test for sensitivity analysis in observational studies. *Biometrics*, 73(2):422–430.
- Ross, A. and Smith, T. (2006). The effect of malaria transmission intensity on neonatal mortality in endemic areas. *The American Journal of Tropical Medicine and Hygiene*, 75(2_suppl):74–81.
- Rubin, D. B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, pages 185–203.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.
- Rubin, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328.
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.
- Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American statistical Association*, 91(434):473–489.
- Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. Cambridge University Press.
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Statistics in Medicine*, 26(1):20–36.
- Sahn, D. E. and Stifel, D. C. (2003). Urban–rural inequality in living standards in africa. *Journal of African Economies*, 12(4):564–597.

- Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15.
- Schieve, L. A., Meikle, S. F., Ferre, C., Peterson, H. B., Jeng, G., and Wilcox, L. S. (2002). Low and very low birth weight in infants conceived with use of assisted reproductive technology. *New England Journal of Medicine*, 346(10):731–737.
- Schmiegelow, C., Matondo, S., Minja, D. T., Resende, M., Pehrson, C., Nielsen, B. B., Olomi, R., Nielsen, M. A., Deloron, P., Salanti, A., Lusingu, J., and Theander, T. G. (2017). Plasmodium falciparum infection early in pregnancy has profound consequences for fetal growth. *The Journal of Infectious Diseases*, 216(12):1601–1610.
- Scott, W. R. (2012). *Group Theory*. Courier Corporation.
- Selvin, S. and Janerich, D. T. (1971). Four factors influencing birth weight. *British Journal of Preventive & Social Medicine*, 25(1):12.
- Shadish, W. R. (2010). Campbell and rubin: A primer and comparison of their approaches to causal inference in field settings. *Psychological Methods*, 15(1):3.
- Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.
- Shauliy-Aharonov, M. (2020). An exact test with high power and robustness to unmeasured confounding effects. *Statistics in Medicine*, 39(8):1041–1053.
- Shepherd, B. E., Gilbert, P. B., Jemai, Y., and Rotnitzky, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to hiv vaccine trials. *Biometrics*, 62(2):332–342.
- Shepherd, B. E., Redman, M. W., and Ankerst, D. P. (2008). Does finasteride affect the severity of prostate cancer? a causal sensitivity analysis. *Journal of the American Statistical Association*, 103(484):1392–1404.
- Shu, D. and Yi, G. Y. (2019). Weighted causal inference methods with mismeasured covariates and misclassified outcomes. *Statistics in Medicine*, 38(10):1835–1854.
- Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., Cercek, A., Smith, R. A., and Jemal, A. (2020). Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(3):145–164.
- Silber, J. H., Rosenbaum, P. R., Ross, R. N., Ludwig, J. M., Wang, W., Niknam, B. A., Hill, A. S., Even-Shoshan, O., Kelz, R. R., and Fleisher, L. A. (2016). Indirect

- standardization matching: assessing specific advantage and risk synergy. *Health Services Research*, 51(6):2330–2357.
- Slepian, D. (1962). The one-sided barrier problem for gaussian noise. *Bell System Technical Journal*, 41(2):463–501.
- Small, D. S., Cheng, J., Halloran, M. E., and Rosenbaum, P. R. (2013). Case definition and design sensitivity. *Journal of the American Statistical Association*, 108(504):1457–1468.
- Smith, D. L., Guerra, C. A., Snow, R. W., and Hay, S. I. (2007). Standardizing estimates of the plasmodium falciparum parasite rate. *Malaria Journal*, 6(1):1–10.
- St.Clair, T. and Cook, T. D. (2015). Difference-in-differences methods in public finance. *National Tax Journal*, 68(2):319–338.
- Strobino, D. M., Ensminger, M. E., Kim, Y. J., and Nanda, J. (1995). Mechanisms for maternal age differences in birth weight. *American Journal of Epidemiology*, 142(5):504–514.
- Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1.
- Stuart, E. A. and Hanna, D. B. (2013). Commentary: Should epidemiologists be more sensitive to design sensitivity? *Epidemiology*, 24(1):88–89.
- Sulyok, M., Rückle, T., Roth, A., Mürbeth, R. E., Chalon, S., Kerr, N., Samec, S. S., Gobeau, N., Calle, C. L., Ibáñez, J., Sulyok, Z., Held, J., Gebru, T., Granados, P., Brückner, S., Nguetse, C., Mengue, J., Lalremruata, A., Sim, B. K. L., Hoffman, S. L., Möhrle, J. J., Kremsner, P. G., and Mordmüller, B. (2017). Dsm265 for plasmodium falciparum chemoprophylaxis: a randomised, double blinded, phase 1 trial with controlled human malaria infection. *The Lancet Infectious Diseases*, 17(6):636–644.
- Thompson, I. M., Goodman, P. J., Tangen, C. M., Lucia, M. S., Miller, G. J., Ford, L. G., Lieber, M. M., Cespedes, R. D., Atkins, J. N., Lippman, S. M., et al. (2003). The influence of finasteride on the development of prostate cancer. *New England Journal of Medicine*, 349(3):215–224.
- Valea, I., Tinto, H., Drabo, M. K., Huybregts, L., Sorgho, H., Ouedraogo, J.-B., Guiguemde, R. T., Van Geertruyden, J. P., Kolsteren, P., D'Alessandro, U., and for the FSP/MISAME study Group (2012). An analysis of timing and frequency of malaria infection during pregnancy in relation to the risk of low birth weight, anaemia and perinatal mortality in burkina faso. *Malaria Journal*, 11(1):71.

- Van de Poel, E., Hosseinpoor, A. R., Jehu-Appiah, C., Vega, J., and Speybroeck, N. (2007). Malnutrition and the disproportional burden on the poor: the case of ghana. *International Journal for Equity in Health*, 6(1):1–12.
- Van der Vaart, A. W. (2000). *Asymptotic Statistics*, volume 3. Cambridge University Press.
- VanderWeele, T. J. and Peng, D. (2007). Sensitivity analysis in observational research: introducing the e-value. *Annals of Internal Medicine*, 6(1):1–12.
- Visconti, G. and Zubizarreta, J. R. (2018). Handling limited overlap in observational studies with cardinality matching. *Observational Studies*, 4:217–249.
- Volpp, K. G., Rosen, A. K., Rosenbaum, P. R., Romano, P. S., Even-Shoshan, O., Canamucio, A., Bellini, L., Behringer, T., and Silber, J. H. (2007). Mortality among patients in va hospitals in the first 2 years following acgme resident duty hour reform. *JAMA*, 298(9):984–992.
- Walker, P. G., ter Kuile, F. O., Garske, T., Menendez, C., and Ghani, A. C. (2014). Estimated risk of placental infection and low birthweight attributable to plasmodium falciparum malaria in africa in 2010: a modelling study. *The Lancet Global Health*, 2(8):e460–e467.
- Walker, S. P., Powell, C. A., Grantham-McGregor, S. M., Himes, J. H., and Chang, S. M. (1991). Nutritional supplementation, psychosocial stimulation, and growth of stunted children: the jamaican study. *The American Journal of Clinical Nutrition*, 54(4):642–648.
- Walsh-Kelly, C. M., Melzer-Lange, M. D., Hennes, H. M., Lye, P., Hegenbarth, M., Sty, J., and Starshak, R. (1995). Clinical impact of radiograph misinterpretation in a pediatric ed and the effect of physician training level. *The American Journal of Emergency Medicine*, 13(3):262–264.
- Weiss, D. J. and Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors*, 45(1):104–116.
- West, S. G. and Thoemmes, F. (2010). Campbell’s and rubin’s perspectives on causal inference. *Psychological Methods*, 15(1):18.
- WHO (1986). Use and interpretation of anthropometric indicators of nutritional status. *Bulletin of the World health organization*, 64(6):929.
- WHO (2006). Who child growth standards: length/height-for-age, weight-for-

- age, weight-for-length, weight-for-height and body mass index-for-age: methods and development.
- WHO (2008a). Global malaria action plan 1 (2000–2015). *Roll Back Malaria Partnership/World Health Organization*.
- WHO (2008b). *Worldwide Prevalence of Anaemia 1993-2005: WHO Global Database on Anaemia*.
- WHO (2016). World malaria report 2016. Technical Report Licence: CC BY-NC-SA 3.0 IGO, Geneva: World Health Organization.
- WHO (2017). Stunting in a nutshell.
- WHO (2019). World malaria report 2019. Technical Report Licence: CC BY-NC-SA 3.0 IGO, Geneva: World Health Organization.
- 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.
- Wittram, C., Maher, M. M., Yoo, A. J., Kalra, M. K., Shepard, J.-A. O., and McCloud, T. C. (2004). Ct angiography of pulmonary embolism: diagnostic criteria and causes of misdiagnosis. *Radiographics*, 24(5):1219–1238.
- Wood, L., Egger, M., Gluud, L. L., Schulz, K. F., Jüni, P., Altman, D. G., Gluud, C., Martin, R. M., Wood, A. J., and Sterne, J. A. (2008). Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ*, 336(7644):601–605.
- Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer.
- Zeka, A., Melly, S. J., and Schwartz, J. (2008). The effects of socioeconomic status and indices of physical environment on reduced birth weight and preterm births in eastern massachusetts. *Environmental Health*, 7(1):60.
- Zhang, B. and Small, D. S. (2020). A calibrated sensitivity analysis for matched observational studies with application to the effect of second-hand smoke exposure on blood lead levels in children. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 69(5):1285–1305.
- Zhang, K., Small, D. S., Lorch, S., Srinivas, S., and Rosenbaum, P. R. (2011). Using split samples and evidence factors in an observational study of neonatal outcomes. *Journal of the American Statistical Association*, 106(494):511–524.

- Zhao, Q. (2018). On sensitivity value of pair-matched observational studies. *Journal of the American Statistical Association*.
- Zhao, Q., Small, D. S., and Rosenbaum, P. R. (2018). Cross-screening in observational studies that test many hypotheses. *Journal of the American Statistical Association*, 113(523):1070–1084.
- Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1997). Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Transactions on Mathematical Software (TOMS)*, 23(4):550–560.
- Zubizarreta, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371.
- Zubizarreta, J. R., Cerda, M., and Rosenbaum, P. R. (2013). Effect of the 2010 Chilean earthquake on posttraumatic stress reducing sensitivity to unmeasured bias through study design. *Epidemiology (Cambridge, Mass.)*, 24(1):79.
- Zubizarreta, J. R., Paredes, R. D., and Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *The Annals of Applied Statistics*, 8(1):204–231.