MODERN OPTIMIZATION IN OBSERVATIONAL STUDIES

Colin B. Fogarty

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2016

Supervisor of Dissertation

Dylan S. Small
Professor of Statistics

Graduate Group Chairperson

Eric T. Bradlow
K.P. Chao Professor, Professor of Marketing, Statistics, and Education

Dissertation Committee

Paul R. Rosenbaum, Robert G. Putzel Professor, Professor of Statistics

Andreas Buja, Liem Sioe Liong/First Pacific Company Professor, Professor of Statistics

MODERN OPTIMIZATION IN OBSERVATIONAL STUDIES

For Beatrice and Janice

ACKNOWLEDGEMENT

I would like to start by thanking my advisor, Dylan. I am simultaneously indebted to and in awe of the care and dedication given by you to each and every one of your students. Time and time again I have been impressed by your seemingly boundless knowledge of the literature, your insights into the benefits and limitations of existing methods, and your unwavering love of scholarship. Your passion for research and advising is nothing short of inspirational, and I am proud to have worked with and learned from you over these past five years.

I would next like to thank my committee members, Andreas and Paul. Andreas, the two classes that I took from you were fundamental in shaping my statistical intuition. In his poem *Maud Muller*, John Greenleaf Whittier writes that "for of all sad words of tongue or pen, The saddest are these: 'It might have been!'" While applicable in most walks of life, I now realize that hope springs eternal when these words are considered in the context of "dataset to dataset" variation and statistical inference. Paul, learning from your writings on observational studies has instilled within me the virtues of clarity, precision and conviction in writing. Each sentence should have a purpose, each theorem a necessity. I am also grateful for your kindness and willingness to meet with me to discuss and share ideas. Given the contents of this dissertation, it goes without saying that your contributions to the field have had a profound impact on my thinking and interests.

I would also like to thank the entirety of the Wharton Statistics Department for creating such a welcoming environment and making my five years as a PhD student so enjoyable. To the many professors with whom I have interacted - thank you for your time, your friendliness and for sharing your insights and perspectives. To our wonderful staff - in short, thank you for making my life so easy, be it through scheduling, reserving rooms, facilitating recommendation letters, help with computing, help with funding, or any of the other myriad ways you go above and beyond. To my cohort, Ville, Kory, Tung, Julie, and Justin - thank you for your friendship, and thank you for your willingness to collaborate as we went through

courses together. I have learned so much from each and every one of you. To the rest of the students with whom I have overlapped - thank you for your camaraderie, for your encouragement, and for your willingness to unwind after periods of hard work.

Thanks and appreciation are, of course, also in order for my family. Thank you so much for your love and encouragement throughout the years. Thank you for providing an environment which fostered independence while making it obvious that help was only a call away. Thank you for always being there for me through times of joy and times of hardship. No matter what life has thrown, and may throw, my way, I know I have and will always have your love and support.

Finally, to my loving wife, Beatrice. You are my inspiration and my motivation. You are the limitless source of positivity that drives me to be the best person I can be. Thank you for everything you do, and for everything you are.

ABSTRACT

MODERN OPTIMIZATION IN OBSERVATIONAL STUDIES

Colin B. Fogarty

Dylan S. Small

Perhaps the best known use of modern techniques for optimization in observational studies is within matching algorithms, wherein treated units are placed into matched sets with similar control units to adjust for overt biases. While the intuitive appeal of matching has been long understood, its ascent in popularity can be attributed in large part to computational advances in network flow optimization. This dissertation explores how modern optimization can be leveraged to address other problems in observational studies. First, we demonstrate how, in the absence of covariate overlap, the *maximal box problem* can be used to define an interpretable study population wherein inference can be conducted without extrapolating on important variables. Next, we discuss how integer programming can be used to perform inference, construct confidence intervals, and provide sensitivity analyses for meaningful causal estimands in matched observational studies when the outcomes of interest are binary. Third, we present a method utilizing convex optimization for conducting a sensitivity analysis when there are multiple outcome variables of interest which, we show, can help attenuate the loss in power from accounting for multiple comparisons when assessing the robustness of a study's findings to unmeasured confounding. Finally, we present methods for conducting a sensitivity analysis for the average treatment effect with continuous outcome variables with and without assuming a known direction of effect.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

In an ideal world there would be no need for observational studies; any hypothesized causal relationship would be tested through controlled randomized experiments, with randomization conferring both a "reasoned basis for inference" (Fisher, 1935) and protection against unmeasured confounding. While ethical and logistical constraints make this experimental ideal impossible to attain, researchers should not be deterred from striving towards it when seeking answers to questions that can only be assessed through observational data.

H.F. Dorn advised that the planner of an observational study should always ask himself the question, "how would the study be conducted if it were possible to do it by controlled experimentation?"(Dorn, 1953, p. 680). The idea that the analysis of observational studies should be made experiment-like was strongly advocated by William Cochran, and has proven profoundly influential not only in how observational studies are planned, but also in how they are analyzed. Matching is one strategy which can be viewed in this light. In an observational study employing matching, treated individuals are placed into matched sets with similar control individuals in an attempt to replicate a block randomized experiment. With the advent of the propensity score (Rosenbaum and Rubin, 1983) and advances in optimization routines for matching (Rosenbaum, 1991; Hansen and Klopfer, 2012), matching has entered mainstream usage. In my research, I have demonstrated that advances in optimization that aided matching's ascent can also be leveraged to address a host of seemingly unrelated issues commonly encountered in the design and analysis of observational studies. Through developing methods for estimation and inference in matched observational studies, I hope to further promote the usefulness of matching in the analysis of non-randomized studies. Not only does matching facilitate estimation of and inference for causal effects assuming no unmeasured confounding, but it also provides a framework for assessing the robustness of a study's conclusions to unmeasured confounding through a sensitivity analysis (Rosenbaum, 2002a, Section 4).

Each of the four chapters within the body of the chapter contains a single, self-contained, paper. In the first paper, we investigate the causal effect of admission to an ICU versus to a hospital ward on 60 day mortality rates for sepsis patients. In conducting this analysis, we encountered a problem common in observational studies: a lack of overlap with respect to important covariates. In this application, a lack of overlap arises because many ICU patients are more severely ill than any hospital ward patient. We cannot possibly infer the effect of the admission decision on mortality for the these patients, as we lack patients admitted to hospital wards with whom their outcomes can be fairly compared. Assessment of causal effects for those individuals would represent an analysis of "extreme counterfactuals," resulting in an extrapolation to which the data cannot honestly attest (King and Zeng, 2006). Rather, inference must be restricted to the area of common support (i.e., those patients who were less gravely ill at presentation). Through the *maximal box problem* (Eckstein et al., 2002), we define a study population by incorporating existing methods for identifying individuals outside the area of common support with respect to important covariates while yielding inclusion criteria which are readily interpretable as intervals of values for these variables. By limiting ourselves to important covariates, we are able to verify the efficacy of our method through the use of visual aids such as scatterplots. We then use matching within this study population to adjust for overt biases for *all* covariates, and we only proceed with inference if the balance between the two groups is deemed acceptable. In this way, practitioners can transparently describe the individuals remaining in the study population, and hence the individuals to whom the resulting statistical analysis applies. This paper is joint work with Mark Mikkelsen, David Gaieski, and Dylan Small, and will appear in the *Journal of the American Statistical Association: Applications and Case Studies*.

The second paper discusses difficulties encountered when using randomization inference and the potential outcomes framework in the analysis of observational studies with binary outcomes. Unlike with continuous outcomes, the only natural causal estimands have corresponding hypothesis tests that are *composite* in nature when the outcome variable of interest is binary. This means that there are many allocations of potential outcomes which yield the

same hypothesized value of the causal estimand. Examples of such estimands are the risk difference, risk ratio, and the effect ratio. To reject a null hypothesis for a causal parameter of this sort, we must reject the null for all allocations of the potential outcomes which satisfy the null. The situation is further complicated when conducting a sensitivity analysis, as inference must also account for the potential existence of unmeasured confounding with a range of impacts on the assignment of interventions. We show that hypothesis testing for a composite null with binary outcomes can be performed by solving an integer linear program under the assumption of no unmeasured confounding. When conducting a sensitivity analysis, an integer quadratic program is required. Under mild assumptions, these optimization problems yield the worst-case p-value within the composite null. We show that our formulation is strong, in that the optimal objective value for our integer program closely approximates that of the corresponding continuous relaxation. This allows hypothesis testing and sensitivity analyses to be conducted efficiently even with large sample sizes and large matched sets. We further demonstrate through a simulation study the importance of a thoughtful formulation in solving large-scale discrete optimization problems. This paper is joint work with Pixu Shi, Mark Mikkelsen, and Dylan Small, and will appear in the *Journal of the American Statistical Association: Theory and Methods*.

In the third paper, we discuss how modern optimization lends support towards demonstrating "multiple operationalism" (Campbell, 1988) in an observational study, wherein one predicts a particular direction of effect for multiple outcome variables under the causal theory in question. This strategy is in line with Fisher's advocating of "elaborate theories" as a means to help bridge the gap between association and causation in an observational study; however, when testing hypotheses on multiple outcomes multiple comparisons must be taken into account. This is true not only when assuming no unmeasured confounding, but also when assessing how robust a study's findings are to unmeasured confounding in the subsequent sensitivity analysis. Concerns over a loss in power may lead practitioners to instead investigate the outcome variable they believe a priori will be most affected by the intervention, thus reducing the extent to which Fisher's advice is followed in practice.

We demonstrate that when performing multiple comparisons in a sensitivity analysis, the loss in power from controlling the familywise error rate can be attenuated. This is because unmeasured confounding cannot have a different impact on the probability of assignment to treatment for a given individual depending on the outcome being analyzed. Existing methods for testing the overall truth of multiple hypotheses allow this to occur by combining the results of sensitivity analyses performed on individual outcomes. By solving a quadratically constrained linear program, we are able to perform a sensitivity analysis while avoiding this logical inconsistency. We show that this allows for uniform improvements in the power of a sensitivity analysis when compared to combining individual sensitivity analyses. This is true not only for testing the overall null across outcomes, but also for testing null hypotheses on specific outcome variables when using certain sequential rejection procedures. We illustrate our method through an example examining the impact of smoking on naphthalene levels in the body. This paper is joint work with Dylan Small, and will appear in the *Journal of the American Statistical Association: Theory and Methods*.

In the fourth paper, we present methods for conducting a sensitivity analysis for perhaps the most common summary measure of a treatment's effect, the *average treatment effect*, with continuous outcome variables, Our analysis follows the standard approach for inference on the average treatment effect in randomized experiments by restricting the set of potential outcomes under consideration to those which satisfy an estimated bound on the variance of the average treatment effect. We show that while the problem could be formulated as a large integer program, a solution can be attained to the problem in its greatest generality in linear time. We further discuss the incorporation of an assumption of a known direction of effect, and how integer programming can be used to conduct a sensitivity analysis in this case. We then compare the sensitivity of inferences to unmeasured confounding under a host of assumptions on the potential outcomes, including the assumption of an additive treatment effect. This work remains in progress and is inspired by recent work of Paul Rosenbaum. As an aside, it goes without saying that “inspired by the work of Paul Rosenbaum" is an accurate descriptor of this dissertation in its entirety.

CHAPTER 2 : Discrete Optimization for Interpretable Study Populations and Randomization Inference in an Observational Study of Severe Sepsis Mortality

Joint work with Mark Mikkelsen, David Gaieski, and Dylan Small

2.1. Introduction

2.1.1. Severe Sepsis Incidence and Mortality

Severe sepsis is a leading cause of morbidity and mortality worldwide. It is defined as a systematic inflammatory response to infection that is accompanied by acute organ dysfunction. Angus et al. (2001) estimate that severe sepsis afflicts roughly 750,000 individuals in the United States per year, of whom an estimated 215,000 perish. Gaieski et al. (2013) note that cases of severe sepsis appear to be on the rise. In a recent study, Liu et al. (2014) found that sepsis contributed to one in every two to three deaths in two complementary hospital cohorts, and suggest that "improved treatment of sepsis (potentially a final hospital pathway for multiple other underlying conditions) could offer meaningful improvements in population mortality."

A critical decision along this pathway is whether to admit a patient to an intensive care unit (ICU), or rather to an appropriate hospital ward. It is estimated that approximately 50 percent of severe sepsis patients in the United States are admitted to an ICU after presentation to an emergency department, with the rest being admitted to a hospital ward (Angus and van der Poll, 2013). Recent evidence suggests that admission to a non-ICU setting may be increasing (Whittaker et al., 2015). Severe sepsis varies in degree of gravity at time of presentation to the emergency department. In general, sicker patients tend to be placed in the ICU, and those exhibiting less severe symptoms are often admitted to the hospital ward. Furthermore, Brun-Buisson et al. (1996) and Rohde et al. (2013) note that there are systematic ways in which the epidemiology, site of infection, and organ dysfunctions

appear to vary between ICU and hospital ward patients.

The existing literature offers contrasting opinions on the optimal process of care for severe sepsis patients. Esteban et al. (2007) argue that there is a large population of patients not admitted to the ICU who could "potentially benefit from more aggressive resuscitation and innovative therapies" that are available in the ICU. They found that severe sepsis patients in hospital wards had a higher estimated mortality rate than those who were admitted to the ICU, although their result was not statistically significant. On the other hand, Levy et al. (2008) found that admission to an ICU covered by intensivists may result in worse health outcomes, in part because patients may receive unnecessary (but potentially harmful) therapies or procedures. It is feasible, then, that certain severe sepsis patients may be better off if they were admitted to the hospital ward, as they would not be subjected to interventions in the ICU that are not warranted given their condition. In keeping with this hypothesis, Sundararajan et al. (2005) found that severe sepsis mortality rates among non-ICU patients were lower than those among ICU patients.

The goal of our analysis is to assess the causal effect of ICU admission versus hospital ward admission on health outcomes. To be precise, we aim to compare the average health outcomes if all individuals were admitted to the ICU with the average outcomes if all patients were admitted to the hospital ward. We use data from a retrospective observational cohort study wherein hospital admissions of individuals with severe sepsis to the Hospital of the University of Pennsylvania between January 2005 and December 2009 were examined; see Whittaker et al. (2015) for further details on the data set. We only consider patients without hemodynamic septic shock (a patient has hemodynamic septic shock if the patient has severe sepsis coupled with hypotension after initial fluid resuscitation) because patients with hemodynamic septic shock are almost exclusively admitted to the ICU (ProCESS Trial, 2014). Investigators identified 1507 remaining individuals with severe sepsis but not hemodynamic septic shock, of whom 695 were admitted to an ICU and 812 were admitted to a hospital ward. Thirty covariates detailing demographic information, comorbidities, emergency

Table 1: Covariate Means and Standard Deviations, Original Population and Study Population for Tier 1 Covariates. The first two columns are the covariate means (standard deviations) in the initial study population, and the last two columns are the covariate means (standard deviations) in the study population defined in Section 2.4.3.

	Original Population		Study Population	
Covariate	ICU	Ward	ICU	Ward
Age	60.1	55.1	60.56	55.88
	(17.4)	(18.4)	(17.1)	(18.3)
Charlson comorbity index	2.52	2.41	2.43	2.48
	(2.81)	(2.64)	(2.70)	(2.65)
Initial serum lactate	4.26	2.56	3.22	2.61
	(2.98)	(1.23)	(1.24)	(0.956)
APACHE II score	17.7	13.6	16.9	13.8
	(6.37)	(5.27)	(5.46)	(4.73)

department process of care, and site of infection were identified by expert consultation as germane to the hospital pathway and to health outcomes. We separated our covariates into three tiers of importance based on an *a priori* assessment (i.e. before examining the data set) of their effect on admission decisions and mortality. Our health outcome is a binary variable that takes on the value 1 if a patient died any time between the date of hospital admission and 60 days after hospital admission. The tier 1 covariates are listed in Table 1 along with their means and standard deviations among ICU and hospital ward patients, while the remaining covariates are summarized in Appendix A.1.

A subgroup of severe sepsis patients who are of particular interest to the critical care community are those with *cryptic septic shock*. These are severe sepsis patients who have normal levels of systolic blood pressure (so do not have hemodynamic septic shock) yet exhibit high levels of initial serum lactate ($\geq$ 4 mmol/L) (Puskarich et al., 2011). Initial serum lactate levels refer to the amount of lactic acid in the blood upon presentation to an emergency department. Initial serum lactate levels have been associated with mortality for severe sepsis patients independent of organ dysfunction, and are therefore thought to be a highly useful biomarker for risk-stratifying patients upon presentation to an emergency department (Mikkelsen et al., 2009). Some believe that cryptic septic shock patients should be classified

as septic shock patients and admitted to an ICU by default, while others suggest that there may be no benefit to such a protocol; see Jones (2011) and Rivers et al. (2011) for both sides of the debate. Hence, in addition to comparing ICU versus hospital ward mortality among all severe sepsis patients without hemodynamic septic shock, we would further like to compare mortality within the subgroup of cryptic septic shock patients, as this subgroup may exhibit mortality outcomes that differ from other severe sepsis patients. While only 10% of patients admitted to the hospital wards had cryptic septic shock in our sample, this number was 44% for patients admitted to the ICU.

2.1.2. From Observational Study to Idealized Experiment

Randomization inference provides an appealing framework even when the data are not the result of a randomized experiment. This is in keeping with the advice of H.F. Dorn, as relayed in Cochran (1965), that "the planner of an observational study should always ask himself the question, 'how would the study be conducted if it were possible to do it by controlled experimentation?'" Through matching on observed covariates, we attempt to mimic a well-balanced randomized experiment. Matching methods encourage researcher blinding, since matched sets can and should be constructed without looking at the outcome of interest. Using randomizations within this idealized experiment as the basis for inference also allows us to assess the robustness of a study's finding to unmeasured confounding through a sensitivity analysis. See Rosenbaum (2002a) for a discussion of using randomization inference within observational studies.

Towards this end, we employ covariate matching to account for measured confounders that may bias our comparison of 60 day mortality rates if all patients had been admitted to the ICU versus if all patients had been admitted to the hospital ward, and then conduct inference with respect to the match that is produced; see Stuart (2010) for a comprehensive overview of common matching algorithms. Full matching, the algorithm used herein, is a type of matching algorithm that optimally assigns individuals into strata consisting of either one treated unit and many control units or one control unit and many treated units,

and is particularly appealing for studies where the ratio of treated individuals to control individuals is close to 1:1. See Rosenbaum (1991) and Hansen (2004) for additional details on full matching.

In Section 2.2, we discuss the randomized experiment that full matching aims to replicate. We begin our analysis in Section 2.3, where we discuss an issue encountered within our comparison of hospital wards and ICU that is common to many observational studies: an inherent lack of covariate overlap. In Section 2.4, we discuss how the *maximal box problem* marries together existing methods for addressing lack of covariate overlap with the intuitive appeal of a study population whose boundaries are clearly defined in terms of important covariates.

Section 2.5 lays out the necessary framework for conducting inference on the *average treatment effect* in the idealized experiment we aim to uncover. Difficulties arise due to the composite nature of a null hypothesis on the average treatment effect, in that different allocations of potential outcomes can yield the same average treatment effect while inducing different randomization distributions for its estimate. We overcome these difficulties by finding a sharp upper bound on the variance of the estimated average treatment effect over all elements of the composite null, which under a normal approximation allows us to carry out inference for the composite null in question. In Section 2.6, we apply our methodology to our sepsis example.

Though seemingly unrelated, our solutions for defining an interpretable study population and conducting randomization inference on the average treatment effect with binary outcomes both utilize methods from discrete optimization. Traditionally, discrete optimization problems were viewed as tractable if the worst-case instance could be solved by an algorithm that grows polynomially in the instance's size, and statisticians have typically limited themselves to using algorithms of this type. Both of the problems we pose are $\mathcal{NP}$-hard, meaning that there is no known polynomial time algorithm for the worst-case instances of these problems. However, there have been recent advances in solving typical cases of these

problems such that a typical case of these problems can often be solved in a reasonable amount of time (Schrijver, 2003). In a recent paper, Zubizarreta (2012) highlighted the usefulness of mixed integer programming for attaining well balanced matched sets. We illustrate that when applying the methods described in this paper to our data set, solutions can be attained in a matter of seconds. Through the methods developed in this work, we hope to further emphasize the usefulness of discrete optimization for observational studies and statistics in general.

2.2. Review of Causal Inference via Matching

2.2.1. Notation For a Stratified Randomized Experiment

Suppose there are I total strata, the i^{th} of which contains $n_i \geq 2$ individuals. In each stratum, $m_i \geq 1$ individuals receive the treatment, $n_i - m_i$ individuals receive the control, and $\min\{m_i, n_i - m_i\} = 1$. Furthermore, m_i is fixed across randomizations, resulting in n_i distinct assignments to treatment and control for the i^{th} stratum. Assignments are independent between distinct strata. Under the potential outcomes framework with binary responses, each individual has two potential binary outcomes: one under treatment, r_{Tij}, and one under control, r_{Cij}, which are 1 if an event would occur and 0 otherwise. The true treatment effect for individual j in stratum i is $\delta_{ij} = r_{Tij} - r_{Cij}$, and is unobservable as each individual receives either treatment or control. The observed response for each individual is $R_{ij} = r_{Tij}Z_{ij} + r_{Cij}(1 - Z_{ij})$, where Z_{ij} is an indicator variable that takes the value 1 if individual j in stratum i is assigned to the treatment; see, for example, Neyman (1923) and Rubin (1974). Each individual has observed covariates $\mathbf{x}_{ij}$.

There are $N = \sum_{i=1}^{I} n_i$ individuals in the study, of whom $N_T = \sum_{i=1}^{I} m_i$ receive the treatment and $N_C = N - N_T$ receive the control. Let $\mathbf{R} = (R_{11}, R_{12}, ..., R_{I,n_I})^T$ and $\mathbf{Z} = (Z_{11}, Z_{12}, ..., Z_{I,n_I})^T$. Let Ω be the set of $\prod_{i=1}^{I} n_i$ possible values $\mathbf{z}$ of $\mathbf{Z}$ under the given stratification. In a randomized experiment, randomness is modeled through the assignment vector; each $\mathbf{z} \in \Omega$ has probability $1/|\Omega|$ of being selected. Hence, quantities dependent

on the assignment vector such as $\mathbf{Z}$ and $\mathbf{R}$ are random, whereas r_{Tij}, r_{Cij}, $\mathbf{x}_{ij}$ are fixed quantities. Let $\mathcal{F} = \{r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, i = 1, .., I, j = 1, ..., n_i\}$. For a randomized experiment, we can then write that $\mathbb{P}(Z_{ij} = 1|\mathcal{F}, \mathbf{Z} \in \Omega) = m_i/n_i$, $i = 1, .., I; j = 1, ..., n_i$ and that $\mathbb{P}(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathbf{Z} \in \Omega) = 1/|\Omega|$.

2.2.2. Matching and Observational Studies

In an observational study, we begin with an unmatched study population of size N. Matching methods aim to create strata where the constituent individuals have similar covariate values, or at a minimum similar probabilities of assignment to treatment (Rosenbaum and Rubin, 1983; Stuart, 2010). Once a match is obtained, the acceptability of the resulting stratification is assessed for covariate balance through the use of various diagnostics, the most common of these being the standardized difference (Rosenbaum, 2010). Let the notation introduced in Section 2.2.1 now apply to the stratification yielded by the matching algorithm. If the match passes the balance diagnostics, randomization inference then proceeds under the assumptions of no unmeasured confounding, common support for the assignment probabilities, and equal probabilities of assignment within a matched set. The assumption of no unmeasured confounding states that given the observed covariates, the probabilities of assignment to treatment are independent of the potential outcomes, that is $\mathbb{P}(Z_{ij} = 1|\mathbf{x}_{ij}) = \mathbb{P}(Z_{ij} = 1|\mathbf{x}_{ij}, r_{Tij}, r_{Cij})$, $i = 1, ..., I; j = 1, ..., n_i$. This probability is known as the *propensity score*, and we denote it by $e(\mathbf{x}_{ij})$. The assumption of common support for the assignment probabilities can be written as $0 < e(\mathbf{x}_{ij}) < 1$, $i = 1, ..., I; j = 1, ..., n_i$. Finally, the assumption of equal probability of treatment assignment within a matched set can be written as $e(\mathbf{x}_{ij}) = e(\mathbf{x}_{ik})$ for all $i = 1, ..., I$; $j, k = 1, ..., n_i$. Under these assumptions, we have that $\mathbb{P}(Z_{ij} = 1|\mathcal{F}, \mathbf{Z} \in \Omega) = m_i/n_i$, $i = 1, .., I; j = 1, ..., n_i$ and that $\mathbb{P}(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathbf{Z} \in \Omega) = 1/|\Omega|$, thus recovering the randomized experiment described in Section 2.2.1.

2.3. Lack of Common Support

2.3.1. Imbalance Caused by Limited Covariate Overlap

We begin by conducting a full match on our entire study population. As was previously noted, we have 30 pre-treatment covariates that were deemed important for both the probability of admission to the ICU versus the ward and for the outcome. Of these, 13 contained missing values; see Appendix A.1 for the percentages of missing observations for these 13 covariates. To account for this, we include 13 new missingness indicators, and fill in the missing values with the mean of the covariates. As is discussed in Rosenbaum and Rubin (1984) and Rosenbaum (2010, Section 9.4), this facilitates balancing both the observed covariates and the pattern of missingness between the two groups being compared. We also include an indicator for whether an individual has cryptic septic shock. We thus have 44 covariates that could be used in constructing our matched sets. In determining which variables to match on, the avoidance of various types of "collider-bias" (Greenland, 2003) must be considered. We first do not control for any post-treatment variables in order to avoid biases that stem from controlling for the consequence of an exposure. One particular type of collider bias, M-bias, can be induced even when only controlling for pre-treatment variables. Despite this, we choose to control for all 44 of these pre-treatment covariates because of the work of Ding and Miratrix (2014), simulation studies of Liu et al. (2012), and arguments of Rubin (2009) that suggest that biases stemming from not controlling for a relevant pre-treatment covariate tend to be more substantial than those that are caused by M-bias.

We use rank-based Mahalanobis distance with a propensity score caliper of 0.2 standard deviations as our distance metric between ICU and hospital ward patients, where the propensity scores are estimated via a logistic regression of our covariates on the treatment indicator; for further discussion on the role of propensity score calipers in multivariate matching, see Rosenbaum (2010, Section 8.3). In addition, we match exactly on the cryptic septic shock indicator, meaning that each stratum produced by the full match must either contain all cryptic septic shock patients or none. We use standardized differences, defined as a weighted

difference in means divided by the pooled standard deviation between groups before matching, to assess balance in our resulting match for the remaining covariates (Stuart and Green, 2008). A common rule of thumb is to deem the balance of a resulting match acceptable if all absolute standardized differences fall below 0.1 (Rosenbaum, 2010). We modify this rule slightly based on our covariate importance tiers, using thresholds of 0.05, 0.10, and 0.15 for the standardized differences of tiers 1, 2, and 3 respectively. Thus, we require more stringent balance for those covariates that are deemed to be of highest importance for the admission decision and for mortality.

We first perform an unrestricted full matching. Without any restrictions, full matching can produce extremely large strata. When applied to our data set, there are strata with ratios of hospital ward patients to ICU patients of 37:1, 1:21, 1:32, and 1:65. Noting the potential for outlandishly large strata, Hansen (2004) advocates placing a bound on the maximal allowable strata size in order to increase the effective sample size (and thus, the power of the resulting analysis). In keeping with this, we also performed full matches with restricted ratios of hospital ward patients to ICU patients within each stratum, with ratios ranging from 2:1, 1:2 to 15:1, 1:15. Neither the unrestricted full match nor any of the restricted full matches resulted in an adequately balanced matched sample based on our standardized difference thresholds.

Our failure to attain a suitably balanced stratification does not suggest a deficiency with full matching; to the contrary, no matching algorithm should be able to produce a suitably balanced stratification without discarding individuals, as there is a severe lack of covariate overlap between patients admitted to the ICU and patients admitted to the hospital wards. Two covariates that were out of balance in all of the restricted ratio matches were initial serum lactate levels and APACHE II scores. As is described in Section 2.1, initial serum lactate is believed to be important for both the admission decision and for health outcomes, while the APACHE II score is a measure of disease severity using physiologic variables and chronic health conditions (Knaus et al., 1985). As Figure 1 displays, virtually all of the

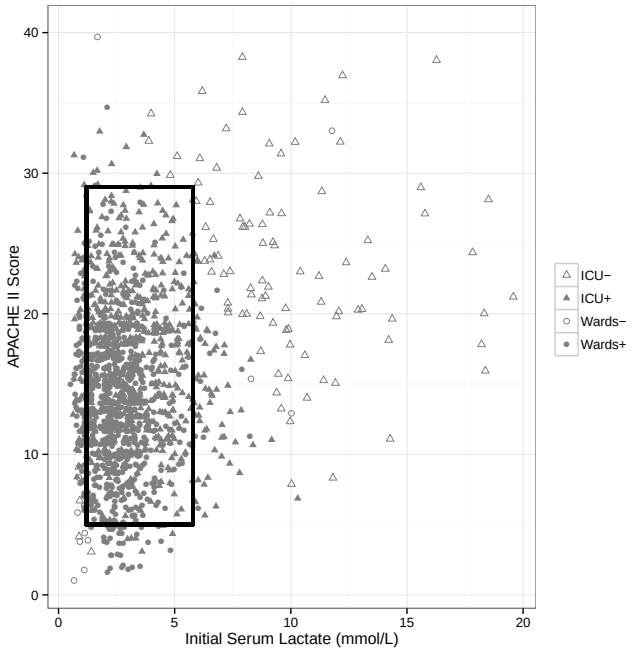

Figure 1: Lack of Common Support and the Maximal Box. This figure shows a scatter plot of initial serum lactate levels and APACHE II scores. The plot also shows the *maximal box*, which is the solution to the optimization problem posed in Section 2.4. The rectangular boundaries represent the study population identified as having a common support, wherein subsequent inference will be restricted. It was formed by finding the rectangle containing the largest number of filled points, subject to excluding all hollow points in the plot. The triangles represent ICU patients, and the circles represent hospital ward patients. Whether a point is filled or hollow is described in Section 2.4.3 in detail, and has to do with whether or not it was determined that a given individual was in the area of viable common support for his or her observed tier 1 covariates. Points are jittered to avoid overplotting.

patients admitted to the hospital ward lie in the lower left hand quadrant of the scatterplot of APACHE II scores versus initial serum lactate levels. Naturally, this lack of overlap arises because many ICU patients are more severely ill than any hospital ward patient. We cannot possibly infer the effect of admission to the ICU versus the hospital ward on mortality for the severely ill ICU patients, as we lack patients admitted to the hospital wards with which the outcomes of those ICU patients can be fairly compared. Assessment of causal effects for those individuals would represent an analysis of "extreme counterfactuals," resulting in an extrapolation to which the data cannot honestly attest (King and Zeng, 2006). Rather, inference about the effect of being admitted to an ICU or a hospital ward on mortality must be restricted to the area of common support (i.e., those patients who were less gravely ill at

presentation), a fact to which restricted ratio full matches bear testament in their inability to attain suitable balance.

2.3.2. Different Types of Overlap

Before proceeding, we discuss a few different notions of covariate overlap. The first notion, which we call strong overlap, is that for every treated unit in the data, there is a control unit that has similar or the same covariate values and that for every control unit, there is a treated unit that has similar or the same covariate values. While strong overlap is most desirable and can be readily diagnosed in low dimensions through visual tools such as scatterplots, it is difficult to obtain when there are a moderate or high number of covariates because of the curse of dimensionality. The second notion, which we call interpolation overlap, is that for any treated unit, an estimate of that treated unit's counterfactual control potential outcome given the unit's covariates can be inferred through an interpolation rather than an extrapolation of the observed control outcomes and that for any control unit, an estimate of that control's unit counterfactual potential outcome can also be inferred through interpolation. King and Zeng (2006) present an operational way to check for interpolation overlap by means of the convex hull of the treated and control covariate distributions. According to their criterion, one is performing interpolation if a given treated (control) individual is in the convex hull of the control (treated) covariate distributions, and is performing extrapolation otherwise. Interpolation overlap then exists if all treated units are in the convex hull of the control units, and all control units are in the convex hull of the treated units. Unfortunately, as noted in King and Zeng (2006) their interpolation overlap criterion is also difficult to satisfy in moderate and high dimensions. In Appendix A.2, we demonstrate through a simulation study that even when the treated and control covariate distributions are identical, the number of individuals for which "interpolation" is identified as being performed by the convex hull diagnostic decreases substantially as the covariate dimension increases.

2.3.3. Existing Methods for Achieving Overlap

A lack of overlap is typically addressed by defining a study population restriction wherein adequate overlap can be attained. Many methods are motivated by the fact that, asymptotically, strong overlap is present if and only if the propensity score at a given covariate value, $e(\mathbf{x}_j)$, is bounded away from 0 and 1 for all individuals $j \in \{1, ...N\}$. In this sense, the propensity score provides a scalar indication of both the existence of and the extent of covariate overlap. Dehejia and Wahba (1999) recommend removing treated units whose propensity scores are larger than the maximal propensity score among the control units, and removing control units whose propensity score are smaller than the minimal propensity score among the treated units. Crump et al. (2009) define a study population by seeking the subset of the covariate space which minimizes the efficiency bound for the variance of the study population average treatment effect. Based on this optimality criterion, they find that for a wide range of distributions a close approximation to the optimal selection rule is to drop all units with estimated propensity scores outside of [0.1, 0.9]. One concern with propensity score approaches for attaining overlap for finite sample inference is that while boundedness away from 0 and 1 implies strong overlap asymptotically, for finite samples treated (control) individuals with nonzero propensity scores may still lack comparable control (treated) individuals in terms of their observed covariates. Another concern is that these propensity scores must be estimated, so that individuals with nonzero estimated propensity scores may nonetheless fall outside the area of overlap.

Other methods directly deal with the covariates themselves when defining a new study population. King and Zeng (2006) identify a multivariate space wherein one performs interpolation rather than extrapolation by removing treated individuals whose covariates lie outside of the convex hull of the covariates for the control individuals, and removing control individuals whose covariates lie outside of the convex hull of the covariates for the treated individuals. Rosenbaum (2012) describes a method for optimal subsampling wherein one chooses an upper bound on how many treated units can be removed from the resulting

matched sample. Hill and Su (2013) employ Bayesian Additive Regression Trees (Chipman et al., 2010) to identify areas of common support, using the fact that the variability of individual-level conditional expectations tend to increase drastically in such areas. Individuals are then classified as being inside or outside the area of common support based on thresholds for these variances.

Though easy to implement and often accompanied by theoretical justifications, the resulting study population returned by these methods is often unappealing as it may be difficult to interpret in terms of the covariates themselves. This makes it difficult to succinctly and transparently describe the individuals to whom the performed inference applies. Furthermore, for study populations defined by propensity scores alone, a researcher's notion of which individuals have high or low "propensity" for treatment may be vastly different from the individuals designated as such through fitting a propensity score model to the data. A practitioner not participating in the study could then have a misconception of the individuals to whom the inference applies based on his or her preconceived notion of which individuals are likely to receive treatment or control. In his *Design of Observational Studies* book, Rosenbaum advises that when excluding extreme individuals "it is usually better to go back to the covariates themselves, $\mathbf{x}_j$, perhaps redefining the population under study to be a subpopulation of the original population" (Rosenbaum, 2010, Section 3.3.3). Stuart (2010) further echoes this sentiment, arguing that "it can help the interpretation of results if it is possible to define the discard rule using one or two covariates" (Stuart, 2010, page 15).

To illustrate the potential confusion arising from a study population definition in terms of propensity scores, suppose we decided to apply the suggestion of Crump et al. (2009) to our tier 1 covariates in order to define our study population. In its most succinct form, the resulting study population would be defined as $\{i : \text{logit}(3.5 - 0.0049(\text{age}_i) + 0.069(\text{CCI}_i) - 0.46(\text{init. ser. lac}_i) - 0.12(\text{APACHE II}_i)) \in [0.1, 0.9]\}$. The boundaries of this set would likely hold little meaning to practitioners, as it is hard to characterize *qualitatively* the

individuals who fall within these bounds. Inference performed on this subset would pertain to a set of individuals who lack a clear characterization on the basis of the covariates of interest themselves, limiting how actionable the findings may be.

Traskin and Small (2011) suggest a tree based approach for defining an internally valid study population based on values of covariates alone. In the first step of their method, the practitioner uses a pre-existing method for study population definition of her choice; any of those described at the beginning of this section would be valid choices. For each individual, this outputs an indicator of whether or not that individual belongs to the area of common support (and hence, should be included in the new study population). The user next fits a regression tree of a designated depth that aims to minimize the probability of misclassification, and defines the study population based on the resulting tree (rather than by the method used in the initial step). While resulting in a markedly more interpretable study population, by their very nature trees result in interval restrictions that are path dependent, rather than intervals that are universally applicable for all individuals.

Restrictions to rectangular regions of the covariate space are appealing as they can be explicitly defined in terms of the intersection of a series of intervals, rather than as a complicated function of the observed covariates. Each interval pertains to a unique covariate, allowing one to paint a coherent description of the resulting study population through covariate-specific constraints. This allows the practitioner to clearly understand the restriction that each covariate imposes on the study population. Currently, little guidance exists on how to define these covariate based inclusion criterion. Ad-hoc choices based on inspection may discard large proportions of individuals, and further may fail to discard individuals who are identified as problematic.

2.3.4. An Attainable Objective

As outlined in this section, there are inherent difficulties with attaining strong overlap in high dimensions. We thus instead seek to define a study population characterized by

three principles which are both attainable and verifiable. Firstly, we would like the study population to demonstrate overlap with respect to those covariates deemed most important for the treatment and the outcome. By focusing on attaining overlap for a small set of important covariates, both strong and interpolation overlap can be potentially obtained for a reasonably sized study population. Furthermore, the overlap with respect to these most important covariates can be verified using visual aids such as scatterplots. Secondly, the resulting study population should be such that balance can be attained on all covariates. As balance is a property of the marginal distributions for the treated and control individuals, standard metrics such as standardized differences can speak to balance being attained for all covariates. Finally, we would like our study population to have a simple definition in terms of important covariates while not being overly wasteful in discarding individuals.

Our approach to achieving these goals is two-fold. We begin by constructing, through the solution to the *maximal box problem*, a study population that incorporates existing methods for identifying individuals outside the area of common support with respect to important covariates, retains as many viable individuals as possible, and is readily interpretable based on important covariates as it defined through the intersection of interval restrictions. After this, we use full matching to arrive upon a stratification that mimics a well-balanced randomized experiment within this study population. We then proceed with inference in the resulting study population only if the balance on all covariates is deemed acceptable.

2.4. Defining a Study Population

2.4.1. The Maximal Box Problem

A box $[\boldsymbol{\ell}, \mathbf{u}]$ is defined to be a closed interval (hyperrectangle) of $\mathbb{R}^p$,

$$[\boldsymbol{\ell}, \mathbf{u}] := \{\mathbf{x} \in \mathbb{R}^p : \ell_i \leq x_i \leq u_i \; \forall i \in \{1, .., p\}\}$$

Suppose one has a finite collection of vectors $\{\mathbf{x}_j\}, j = 1, ..., N$, that can be partitioned

into two disjoint sets of "positive" points, $\mathcal{X}^+$ and "negative" points, $\mathcal{X}^-$. The maximal box problem aims to find the lower and upper boundaries of a box, $[\tilde{\boldsymbol{\ell}}, \tilde{\mathbf{u}}]$, such that the corresponding box contains the maximal number of points in $\mathcal{X}^+$ while containing none of the points in $\mathcal{X}^-$. Explicitly, $[\tilde{\boldsymbol{\ell}}, \tilde{\mathbf{u}}]$ is the arg max of the following optimization problem (MB, for maximal box):

$$\begin{aligned} \text{maximize} \quad & |[\boldsymbol{\ell}, \mathbf{u}] \cap \mathcal{X}^+| \\ \text{subject to} \quad & |[\boldsymbol{\ell}, \mathbf{u}] \cap \mathcal{X}^-| = 0, \end{aligned} \tag{MB}$$

where the notation $|A|$ denotes the number of elements of set A. Henceforth, we refer to $|[\boldsymbol{\ell}, \mathbf{u}] \cap \mathcal{X}^+|$ as the *cardinality* of a box $[\boldsymbol{\ell}, \mathbf{u}]$.

Eckstein et al. (2002) describe the problem in detail. They prove that the problem is $\mathcal{NP}$-hard in general, but is polynomial time for any fixed dimension p. They provide an efficient branch and bound algorithm for solving it, which they show to have modest computation time in practice. They also provide a mixed integer programming formulation of the problem, which facilitates its use with freely available and commercial solvers.

2.4.2. From Maximal Boxes to Study Populations

Let $\mathbf{D}(\mathbf{x}_j, \mathbf{X}, \mathbf{Z})$ be a binary decision rule that determines whether or not a point $\mathbf{x}_j$ needs to be excluded from the analysis to ensure covariate overlap (1 if not, 0 if so). For example, the recommendations of Dehejia and Wahba (1999), $\mathbf{D}_{DW}(\mathbf{x}_j, \mathbf{X}, \mathbf{Z})$, and the rule proposed in Crump et al. (2009) (the simplified version of the rule), $\mathbf{D}_C(\mathbf{x}_j, \mathbf{X}, \mathbf{Z})$, can be written in this form as:

$$\mathbf{D}_{DW}(\mathbf{x}_j, \mathbf{X}, \mathbf{Z}) = \begin{cases} \mathbb{1}\left\{\hat{e}(\mathbf{x}_j) \leq \max\{\hat{e}(\mathbf{x}_k) \text{ s.t. } Z_k = 0\}\right\} & \text{if } Z_j = 1 \\ \mathbb{1}\left\{\hat{e}(\mathbf{x}_j) \geq \min\{\hat{e}(\mathbf{x}_k) \text{ s.t. } Z_k = 1\}\right\} & \text{if } Z_j = 0 \end{cases}$$

$$\mathbf{D}_C(\mathbf{x}_j, \mathbf{X}, \mathbf{Z}) = \mathbb{1}\left\{\hat{e}(\mathbf{x}_j) \in [0.1, 0.9]\right\}$$

Our sets of positive and negative points are then defined based on the selected decision rule, with $\mathcal{X}^+ := \{\mathbf{x}_j : \mathbf{D}(\mathbf{x}_j, \mathbf{X}, \mathbf{Z}) = 1\}$, and $\mathcal{X}^- := \{\mathbf{x}_j : \mathbf{D}(\mathbf{x}_j, \mathbf{X}, \mathbf{Z}) = 0\}$. We then solve (MB) using these designations of positive and negative points. The resulting maximal box is one that contains the largest possible number of observations who could feasibly have been in the study population, while eliminating all individuals who were designated for exclusion. The study population defined by the maximal box has a clear interpretation in terms of the covariates themselves: an individual is in the study population if $\tilde{\boldsymbol{\ell}} \leq \mathbf{x}_j \leq \tilde{\mathbf{u}}$, and is excluded otherwise.

We note that as p (the number of covariates used to define the maximal box) increases, the number of positive points in corresponding maximal box is non-decreasing. At the same time, this increases the potential computational burden, as there are at most $|\mathcal{X}^+|^{2p}$ possible candidates for the boundaries of the maximal box (Eckstein et al., 2002). Thus, in practice we recommend forming the boundaries on the maximal box based on values of the most important covariates. Note that defining a study population on the basis of important covariates can also be justified on the basis of interpretability. If one defined a study population using a maximal box formed from a large number of covariates, the resulting study population would likely be just as cryptic as one determined solely by the estimated propensity scores. Further, Hill and Su (2013) argue that methods for common support restriction should primarily consider those covariates that are most important for the outcome. As such, we seek to define a study population based on the most important pre-treatment covariates. We also recommend using covariates that are not binary for constructing the maximal box as the resulting restriction would either eliminate one of the categories entirely, or (more commonly) be the whole range [0,1]. If there is a binary covariate of considerable importance, we recommend accounting for it by either exactly matching or almost exactly matching on the binary covariate; see Rosenbaum (2010, Sections 9.1 and 9.2) for details.

There is a possibility that the resulting maximal box only contains a small fraction of the positive points. This means there is no easy way to define a region of good overlap between

the treated and control individuals without eliminating the vast majority of the data. In Appendix A.3, we discuss an extension of the maximal box problem posed in Eckstein et al. (2002) that may be appropriate in this setting. This generalization allows for a small number, C, of points marked for exclusion (negative points) to be included within the bounds of the maximal box, which would in turn allow for the incorporation of more positive points; see Appendix A.3 for more discussion on the ramifications of choosing $C > 0$. In our example, we proceed with $C = 0$, thus requiring the exclusion of all points marked as being outside the area of viable support.

2.4.3. Application to Our Original Population

As defining a maximal box with all 44 covariates would yield a highly unwieldy 44 dimensional box with limited interpretability, we instead aim to construct a maximal box using our four tier 1 covariates: age, Charlson comorbidity index, APACHE II scores, and initial serum lactate levels. Our approach is to fit a propensity score model using a logistic regression on our four tier 1 covariates, and to then employ the simplified criterion of Crump et al. (2009) with these propensity scores to determine which observations had to be removed. We use this reduced propensity score model because individuals within the area of common support on our important variables may be nonetheless extreme with respect to other, less important, covariates, which may in turn lead to them being marked for removal if we used the full propensity score model. As our focus is on attaining covariate overlap *and* balance for our most important variables while seeking balance on all other variables, we wanted our exclusion metric to reflect lying in the area of covariate overlap with respect to our most important variables. See Appendix A.4 for a more detailed discussion of this goal and the behavior of alternative strategies. Denoting the tier 1 covariate for individual j as $\mathbf{x}_j^{(1)}$, our decision rule is $\mathbf{D}_{C,\text{Tier1}}(\mathbf{x}_j, \mathbf{X}, \mathbf{Z}) = \mathbb{1}\left\{\hat{e}\left(\mathbf{x}_j^{(1)}\right) \in [0.1, 0.9]\right\}$. This results in 108 individuals being marked for exclusion. We have implemented the branch and bound algorithm of Eckstein et al. (2002) in the `R` programming language (R Development Core Team, 2014), and used it to find our study population; a script for our implementation is provided in the

supplementary materials. For this data set, our implementation took 2 seconds to run on a desktop computer with a 3.40 GHz processor and 16.0 GB RAM.

We created a maximal box using all four tier 1 covariates, and also created one using only initial serum lactate and APACHE II scores. The cardinalities of these boxes were very close to one another (1214 and 1208 respectively). As such, we use the box defined using only initial serum lactate and APACHE II scores for enhanced interpretability. The resulting maximal box is displayed as the rectangle in Figure 1. As can be seen, the study population under investigation can be explicitly defined as those individuals in our initial study whose APACHE II scores are between 5 and 29 and whose initial serum lactate levels are between 1.2 and 5.8 mmol/L. Our study population thus restricts analysis to those individuals who had less severe, but not the least severe, conditions upon presentation to the emergency department. The study population defined by the maximal box includes 701 out of 812 patients admitted into the wards and 507 out of 695 patients admitted to the ICU, resulting in 1208 out of the original 1507 individuals being available for further study; furthermore, it contains 86.3% of all individuals whose estimated propensity scores were deemed acceptable by our decision rule. Table 1 shows the means and standard deviations of the tier 1 covariates among this study population; values for the other covariates can be found in Appendix A.1. As can be seen, restricting ourselves to this study population improved pre-matching balance for many of the covariates.

We now proceed with a full matching on our study population of 1208 individuals whose condition upon presentation was less severe. In so doing, we follow a procedure analogous to the one described in Section 3.1 within our newly defined study population. We first refit our propensity score model using all 44 covariates for the 1208 individuals in our newly defined study population to exploit the so-called balancing property of the propensity score within our population of interest (Rosenbaum and Rubin, 1983). We use rank-based Mahalanobis distance based on all 44 covariates with a propensity score caliper of 0.2 standard deviations computed only for patients in our study population to define distance between ICU and

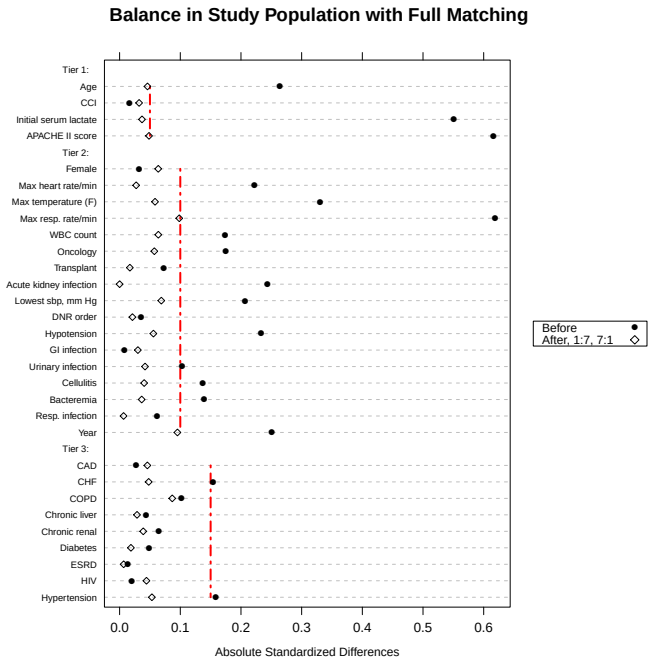

Figure 2: Covariate Imbalances Before and After Full Matching, Study Population. The dotplot (a Love plot) shows the absolute standardized differences without matching, and after conducting a restricted 1:7, 7:1 full matching on our study population. The vertical dotted lines correspond to the standardized difference tolerances for each of the three covariate tiers. Although not shown here, all standardized differences corresponding to our 13 missingness indicators had standardized differences below 0.1, indicating that the pattern of missing data was also balanced between the ICU and hospital ward groups.

hospital ward patients. Further, we require exact matching for the cryptic septic shock indicator. Using the resulting distance matrix, we run a series of full matches with ratios of 1:k, k:1, starting with $k = 2$ and increasing k until a suitably balanced matched sample could be attained. We found that a 1:7, 7:1 full match was able to adhere to the standardized difference tolerances defined in Section 2.3.1, as is displayed in Figure 2.

At this point we have obtained a matched data set that is easy to characterize in terms of bounds on two important covariates and demonstrates balance on all of our covariates. Moving forward, we will treat this match as though it were instead a block randomized experiment with strata of maximal size 8, where in each stratum either one unit is randomly assigned the treatment and the rest receive the control, or one unit is randomly assigned the control and the rest receive the treatment. Using randomizations within this idealized

experiment as the basis for statistical inference, our goal is to assess not only whether there is a substantial difference in mortality rates depending on admission to the ICU or the hospital ward, but also to measure the extent of the effect. In order to do so, we now discuss performing inference and constructing confidence intervals for the average treatment effect in our idealized experiment.

2.5. Randomization Inference for the Average Treatment Effect with Binary Outcomes

The average treatment effect with binary outcomes (also known as the causal risk difference) is the difference between the proportion of positive responses among the potential outcomes under treatment and the potential outcomes under control, $\delta := (1/N)\sum_{i=1}^{I}\sum_{j=1}^{n_i}\delta_{ij}$. It it identifiable under the assumption of strong ignorability (Rosenbaum and Rubin, 1983), and an unbiased estimator of the average treatment effect under a stratified design is given by $\hat{\delta} := \sum_{i=1}^{I}(n_i/N)\hat{\delta}_i$, where $\hat{\delta}_i = \sum_{j=1}^{n_i}(Z_{ij}R_{ij}/m_i - (1-Z_{ij})R_{ij}/(n_i - m_i))$ is the estimated average treatment effect within stratum i (Rosenbaum, 2002a, Section 2.5).

We consider tests of the null hypothesis that $(1/N)\sum_{i=1}^{I}\sum_{j=1}^{n_i}\delta_{ij} = \delta_0$, $\delta_0 \in \{d/N : d \in [-N, N] \cap \mathbb{Z}\}$, where $\mathbb{Z}$ denotes the set of all integers. In reality, a null hypothesis of this form is a large collection of hypotheses on the set of treatment effects, $\boldsymbol{\delta} = [\delta_{11}, \delta_{12}, ..., \delta_{I,n_I}]$. Let $\mathcal{D}_{\delta_0}$ be the set of all $\boldsymbol{\delta}$ such that $(1/N)\sum_{i=1}^{I}\sum_{j=1}^{n_i}\delta_{ij} = \delta_0$ and such that the treatment effects are compatible with the observed data. The latter requirement means that if unit j in stratum i received the treatment in the observed experiment, the value of r_{Tij} is fixed at R_{ij} and hence δ_{ij} can equal either R_{ij} or $R_{ij} - 1$. If said unit received the control, the value of r_{Cij} is fixed at R_{ij}, and δ_{ij} can equal either $-R_{ij}$ or $1 - R_{ij}$. To reject a null hypothesis $(1/N)\sum_{i=1}^{I}\sum_{j=1}^{n_i}\delta_{ij} = \delta_0$, we require that we reject the null hypothesis that the allocation of treatment effects equals $\boldsymbol{\delta}$ for all $\boldsymbol{\delta} \in \mathcal{D}_{\delta_0}$.

2.5.1. Existing Methods

Inspired by the work of Neyman (1923), randomization inference for the average treatment effect is typically conducted by finding a consistent estimator of an upper bound on the variance of the estimated ATE resulting in randomization inference that asymptotically has the proper Type I error rate; see Ding (2014) among many. Robins (1988) improves upon the upper bound of Neyman (1923) for binary outcomes under an unstratified design and uses the resulting upper bound to create confidence intervals that are narrower than those based on a Wald-type procedure. More recently, Aronow et al. (2014) provide asymptotically sharp upper bounds on $\text{var}(\hat{\delta})$ under general potential outcomes.

For a stratified design, the variance for the estimated ATE is

$$\text{var}(\hat{\delta}) = \sum_{i=1}^{I} \frac{n_i^2}{N^2} \left(\frac{S_{Ti}^2}{m_i} + \frac{S_{Ci}^2}{n_i - m_i} - \frac{S_{\delta i}^2}{n_i} \right) \tag{2.1}$$

where $S_{Ti}^2 = \sum_{j=1}^{n_i} (r_{Tij} - \bar{r}_{Ti})^2/(n_i - 1)$, $S_{Ci}^2 = \sum_{j=1}^{n_i} (r_{Cij} - \bar{r}_{Ci})^2/(n_i - 1)$, and $S_{\delta i}^2 = \sum_{j=1}^{n_i} (\delta_{ij} - \bar{\delta}_i)^2/(n_i - 1)$. The procedures of Neyman (1923), Robins (1988) and Aronow et al. (2014) can be readily extended to stratified designs where m_i and $n_i - m_i$ are sufficiently large for each stratum i. However, these procedures have deficiencies when there are strata for which either m_i or $n_i - m_i = 1$, as these procedures require an estimate of the variance of the treated and control groups in each stratum. When m_i or $n_i - m_i = 1$, unbiased estimators for S_{Ti}^2 or S_{Ci}^2 do not exist. Matched sets returned by pair matching, fixed ratio matching, variable ratio matching and full matching have this property, rendering the existing bounding techniques based solely on in-sample estimates inapplicable.

Rigdon and Hudgens (2014) present two methods for conducting randomization inference and constructing confidence intervals for the average treatment effect with binary outcomes in an unstratified design. The first method proceeds by combining two tests on the *attributable effect* of Rosenbaum (2001) and Rosenbaum (2002b) through means of a Bonferroni correction. They then mention that this approach, while potentially conservative, can be

readily applied to stratified randomized experiments. In the second method, hypothesis testing proceeds by conducting randomization inference on δ for all $\boldsymbol{\delta} \in \mathcal{D}_{\delta_0}$, meaning that this procedure has level α for testing the corresponding composite null. Confidence intervals are then constructed by inverting tests for values under the null $\delta_0 \in \{d/N : d \in [-N, N] \cap \mathbb{Z}\}$, where $\mathbb{Z}$ again denotes the set of all integers. In their description, inference is conducted by explicitly performing a randomization test for each $\boldsymbol{\delta} \in \mathcal{D}_{\delta_0}$. Noting the inherent computational burden in this process as N increases in an unstratified experiment, they suggest a Monte Carlo procedure to approximate the required permutation test. For stratified experiments, they suggest that this approach becomes computationally unwieldy quite quickly, thus advocating the use of a potentially conservative method based on the attributable effect in this setting.

Our procedure combines elements of the classical Neyman approach and the hypothesis test inversion approach of Rigdon and Hudgens (2014). Our approach is not purely Neymanian in that although we are testing Neyman's null hypothesis, we do not proceed by seeking a consistent upper bound on $\text{var}(\hat{\delta})$; rather, we explicitly compute the largest value of $\text{var}(\hat{\delta})$ possible among the elements of $\mathcal{D}_{\delta_0}$ for each null hypothesis. The resulting bound on the variance of the average treatment effect for a given null hypothesis is sharp, as it is attained by a member of the composite null $\boldsymbol{\delta}^* \in \mathcal{D}_{\delta_0}$. As a test of a composite null hypothesis is size α only if the supremum over all elements of the composite null of the probability of rejection is α, asymptotically our testing procedure has size α so long as a normal approximation is justified. This is because since the numerator is the same for the test statistic for any null in $\mathcal{D}_{\delta_0}$, namely $\hat{\delta} - \delta_0$, the p-value computed under a normal approximation will be maximized by the member of the composite null with the largest denominator of the test statistic, i.e, the member with the largest variance. Rejection on the basis of this worst-case p-value then implies rejection for all elements of the composite null. For finite samples, discrepancies in actual versus advertised size stem only from the strength of the normal approximation. We show in Appendix A.5 that for our case study, the true distribution of the average treatment corresponding to the worst-case allocations of potential outcomes is well approximated by a

normal distribution. In Appendix A.6, we discuss why our standard errors are necessarily larger than those attained in other testing scenarios (for example, in testing Fisher's sharp null).

As will be discussed in Section 2.5.2, the use of a normal approximation allows us to overcome the computational issues encountered in Rigdon and Hudgens (2014). This normal approximation can be justified under very mild conditions. Let $\sigma_i^2 = n_i^2(S_{Ti}^2/m_i + S_{Ci}^2/(n_i - m_i) - S_{\delta i}^2/n_i)$ be the contribution to $\text{var}(N\hat{\delta})$ from stratum i (i.e., $\sum_{i=1}^I \sigma_i^2/N^2 = \text{var}(\hat{\delta})$), and let $\Sigma = \sum_{i=1}^I \sigma_i^2$. Let n^* be an upper bound on the maximal size of a stratum.

Theorem 1. *If $\Sigma \to \infty$ as $I \to \infty$, then $(N\hat{\delta} - N\delta)/\sqrt{\Sigma} \xrightarrow{d} \mathcal{N}(0,1)$.*

Proof. Since our outcomes are binary, the maximal contribution of an individual summand $n_i\hat{\delta}_i$ to $\sum_{i=1}^I n_i\hat{\delta}_i = N\hat{\delta}$ is bounded in absolute value by n^*. Using Lyapunov's central limit theorem applied to a sequence of independent bounded random variables (Lehmann, 2004, Corrolary 2.7.1), we have that $(N\hat{\delta} - N\delta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ as $I \to \infty$ provided that $\Sigma \to \infty$ as $I \to \infty$. □

This requirement precludes a certain type of degeneracy. Namely, it cannot be the case that only finitely many strata have nonzero variances for $\hat{\delta}_i$. This, coupled with a bound on the maximal stratum size, suffices for asymptotic normality to hold.

2.5.2. Integer Programming for the Maximal Variance

In theory, the maximal variance for a given composite null, $H_0 : \delta = \delta_0$, could be found by enumerating all 2^N possible allocations of unobserved binary potential outcomes, computing $\text{var}(\hat{\delta})$ through (2.1) for each allocation, and finding the maximal variance among the allocations that satisfy $\boldsymbol{\delta} \in \mathcal{D}_{\delta_0}$. Such a naïve approach quickly becomes computationally infeasible: in our application, this would would require enumerating 2^{1208} sets of potential outcomes.

Our approach is to instead pose the problem of maximizing the variance within a composite

null as an integer program. Roughly stated, the resulting integer program optimizes the variance over the values of the unobserved potential outcomes, subject to the resulting allocation of potential outcomes being a member of the composite null. Though many equivalent formulations of the desired optimization problem are possible, the one we choose explicitly avoids symmetric solutions, known to cripple the computation time of integer programs (Margot, 2010), by having each decision variable correspond to a unique distribution on the contribution to the overall estimated average treatment effect from a given stratum. Our approach exploits three essential facts. Firstly, there is often symmetry *between* strata in that (a) $n_i = n_{i'}$, (b) $\sum_{j=1}^{n_i} Z_{ij}R_{ij} = \sum_{j=1}^{n'_i} Z_{i'j}R_{i'j}$ and (c) $\sum_{j=1}^{n_i}(1-Z_{ij})R_{ij} = \sum_{j=1}^{n'_i}(1-Z_{i'j})R_{i'j}$ for symmetric strata i and i', meaning that any allocation of potential outcomes for stratum i is also a feasible allocation for stratum i'. Secondly, there is often symmetry *within* strata in that $Z_{ij} = Z_{ik}$ and $R_{ij} = R_{ik}$ for symmetric individuals j and k in stratum i, meaning that the $\text{var}(\hat{\delta}_i)$ remains the same if the values for the unobserved potential outcome are permuted among symmetric individuals in stratum i. Finally, there is independence between strata which allows us to sum stratum-wise variance contributions together to arrive at the overall variance of the estimated average treatment effect. In combination, these three facts allow this seemingly daunting optimization problem to be solved in a matter of seconds. See Appendix A.7 for a detailed discussion of our integer programming formulation.

2.6. Inference for Severe Sepsis Mortality

We now proceed with randomization inference on the study population defined by our maximal box in Section 2.4.3. As a reminder, this consists of severe sepsis patients without hemodynamic septic shock, with initial serum lactate between 1.2 and 5.8 mmol/L, and with APACHE II scores between 5 and 29. Of the 1208 patients in our study population, 701 were admitted to the hospital ward and 507 were admitted to the ICU. Our causal estimand is the difference between 60 day mortality rates if all patients had been admitted to the ICU and if all patients had been admitted to the hospital ward. Before matching, the unadjusted (and hence potentially biased) estimates for these rates under ICU and hos-

Table 2: Estimated differences in severe sepsis mortality between patients admitted to the ICU and the hospital ward in our study population, both overall and among patients with cryptic septic shock. Positive values favor hospital ward admission, and negative values favor ICU admission. The standard errors reported are the Wald-estimates (i.e. the maximal standard errors at the estimated average treatment effects) and confidence intervals were constructed by inverting a series of tests as described in Section 2.5.1.

	Overall	Cryptic S.S.
Estimated ATE	4.3%	-0.8%
(SE)	(3.7%)	(9.0%)
95% Conf. Int.	[-3.0%; 11%]	[-18%; 17%]

pital ward admissions are 24.3% and 12.0% respectively overall, and are 27.7% and 21.2% respectively within the cryptic septic shock subgroup.

After adjusting for measured confounders through covariate matching, the estimated mortality rates under ICU and hospital ward admission are 19.4% and 15.1% respectively overall, and are 26.0% and 26.8% respectively within the cryptic septic shock subgroup. Table 2 shows the estimated average treatment effects (the differences between proportions under ICU and hospital ward admission) both in our overall study population and among the cryptic septic shock subgroup. We also report 95% confidence intervals, which were formed by inverting a series of hypothesis tests as discussed in Section 2.5.1. Both of these confidence intervals contain 0, indicating that we lack substantial evidence to suggest that there is a nonzero effect both overall and in the cryptic septic shock subgroup. Through our implementation, we were able to construct the reported confidence intervals in 0.42 seconds using `Gurobi` (free for academic use), and 0.72 seconds to solve using `lpSolve` (free for all users) on a desktop computer with a 3.40 GHz processor and 16.0 GB RAM. This demonstrates that confidence intervals can be constructed using our integer programming formulation efficiently using both commercial and freely available solvers.

2.7. Discussion

As expected, we found that common support was not present for the most severely ill sepsis patients. The subset of septic shock patients, which include those with hemodynamic

compromise or evidence of hypoperfusion, are routinely admitted directly to the ICU and therefore an observational study cannot address the effect of these triage decisions. For the population with substantial common support, our findings suggested that there was no clear benefit to direct ICU admission for non-shock, severe sepsis patients. In fact, recognizing our wide confidence intervals, the magnitude of the potential benefit of direct ICU admission after adjusting for all measured confounders through matching at the leftmost extreme of our confidence interval was relatively small at 3%. While larger studies are required to substantiate our findings, under the assumption of no unmeasured confounding our analysis finds no evidence to suggest that the common practice in hospitals with strained ICUs (occupancy rates approaching 100%) to defer ICU admission for many severe sepsis patients results in demonstrable harm for those who are less severely ill at the time of presentation to the emergency department.

By using the maximal box problem to define a study population for further analysis, we arrived at a study population with a readily interpretable definition in terms of important covariates wherein acceptable balance could be attained. One downside of our method is that it is not guaranteed that suitable balance can be attained in the resulting study population. That is, one may arrive at a study population defined in terms of important covariates where it is difficult to find a matching procedure that attains suitable balance on all covariates. One option is to simply iterate: covariates for which suitable balance cannot be achieved can be used in defining a study population through the maximal box problem, and then one could again try to attain balance within the proposed study population. An interesting area for future research would be to create a procedure where the returned study population is guaranteed to have a match with acceptable balance. With fixed ratio matching, recent work on mixed-integer programming matching (Zubizarreta, 2012) and cardinality-matching (Zubizarreta et al., 2014) may provide insight into how to incorporate the balancing constraints into the optimization problem.

In our application, we determined which covariates were most important for the treatment

and the outcome (and hence those for which we seek verifiable overlap) through consultation with subject matter experts. In other applications, practitioners may not want to rely solely on prior information for determining which covariates are important and rather allow the data itself to attest to this. While model selection for the propensity score model can be conducted without concern, one must be careful when assessing the impact of covariates on the outcome variable as it could potentially bias the resulting inference by compromising the "researcher blinding" that makes matching so appealing (Rubin and Waterman, 2006). One path forward would be to employ sample splitting, thus assessing importance of covariates for the outcome using data that is not involved in the matched analysis.

Through our analysis of the impact of ward versus ICU admission on 60 day mortality rates, we have shown that the applicability of discrete optimization in causal inference extends far beyond matching algorithms. In fact, discrete optimization provides a powerful set of tools for solving many problems common to observational studies and, more broadly, statistics in general. The availability of efficient solvers can serve as the impetus for new methods that trade potentially unverifiable model assumptions for an increase in computation time. This is not to say that computational burden should not be considered when developing statistical methodology; rather, it is to caution against limiting the imagination solely on the basis of the computational power of the present day. As history has borne out, what is intractable today may be feasible tomorrow.

CHAPTER 3 : Randomization Inference and Sensitivity Analysis for Composite Null Hypotheses with Binary Outcomes in Matched Observational Studies

Joint work with Pixu Shi, Mark Mikkelsen, and Dylan Small

3.1. Introduction

3.1.1. Challenges for Matched Observational Studies with Binary Outcomes

Matching is a simple, transparent and convincing way to adjust for overt biases in an observational study. In a study employing matching, treated subjects are placed into strata with control subjects on the basis of their observed covariates. In each stratum, there is either one treated unit and one or more similar control units, or one control unit and one or more similar treated units (Hansen, 2004; Rosenbaum, 2010; Stuart, 2010). The overall covariate balance between the two groups is then assessed with respect to the produced stratification, and inference is only allowed to proceed if the balance is deemed acceptable. This procedure encourages researcher blinding, as both the construction of matched sets and the assessment of balance proceed without ever looking at the outcome of interest just as they would in a blocked randomized trial.

Despite our best efforts, observational data can never achieve their randomized experimental ideal as the assignment of interventions was conducted outside of the researcher's control. Nonetheless, randomization inference provides an appealing framework within which to operate for matched observational studies. The analysis initially proceeds as though the data arose from a blocked randomized experiment, with the strata constructed through matching now regarded as existing before random assignment occurred. Randomization inference uses only the assumption of random assignment of interventions to provide a "reasoned basis for inference" in a randomized study (Fisher, 1935). In the associated sensitivity analysis for an observational study, departures from random assignment of treatment within each block due

to unmeasured confounders are considered. The sensitivity analysis forces the practitioner to explicitly acknowledge greater uncertainty about causal effects than would be present in a randomized experiment due to the possibility that unmeasured confounders affect treatment assignment and the outcome (Rosenbaum, 2002a, Section 4).

With binary outcomes, randomization inference and sensitivity analyses in matched observational studies raise computational challenges that have heretofore limited their use. When the outcome is continuous rather than binary and an additive treatment effect is plausible, hypothesis testing and sensitivity analyses for the treatment effect can be conducted for a *simple* null hypothesis, and confidence intervals can then be found by inverting a series of such tests. This is a straightforward task, since the potential outcomes under treatment and control for each individual are uniquely determined by the hypothesized treatment effect (Hodges and Lehmann, 1963). Inference under no unmeasured confounding merely requires a simple randomization test, and a sensitivity analysis can be performed with ease through the asymptotically separable algorithm of Gastwirth et al. (2000). When dealing with binary responses, however, an additive treatment effect model is inapplicable: if an effect exists it is most likely heterogeneous, as the intervention may cause an event for one individual while not causing the event for another. As such, confidence intervals are instead constructed for causal estimands whose corresponding hypothesis tests are *composite* in nature, meaning there are many allocations of potential outcomes which yield the same hypothesized value of the causal estimand; see Rosenbaum (2001, 2002b) for further discussion. To reject a null hypothesis for a causal parameter of this sort, we must reject the null for all values of the potential outcomes which satisfy the null. The situation is further complicated when conducting a sensitivity analysis, as inference must also account for the existence of an unmeasured confounder with a range of impacts on the assignment of interventions within a matched set. We now illustrate these points by investigating the causal effect of one post-hospitalization protocol versus another after an acute care stay on hospital readmission rates.

3.1.2. Motivating Example: Effect of Post-Acute Care Protocols on Hospital Readmission

At the time of discharge after an acute care hospitalization, a fundamental question arises: to where should the patient be discharged? The long-term goal shared by providers and patients envisions a transition home and a return to normalcy, yet a premature discharge home without appropriate guidance could impede a durable recovery.

An important measure of whether a patient has achieved a durable recovery is whether the patient does not need to be readmitted to the hospital within a certain period of time. Different avenues for reducing rehospitalization rates have recently garnered significant attention nationwide (Jencks et al., 2009), and post-acute care is one mechanism through which hospital readmission rates may be improved (Ottenbacher et al., 2014). For individuals who are not gravely ill, post-acute care entails more intensive discharge options than a simple discharge home without further supervision such as discharge home while receiving visits from skilled nurses, physical therapy, and other additional health benefits (referred to henceforth as "home with home health services"); or discharge to an acute rehabilitation center. Post-acute care use is on the rise in the United States; however, post-acute care services can be quite costly, sometimes even rivaling the cost of a hospital readmission (Mechanic, 2014). It is thus of interest to assess the relative merits of various post-acute care protocols for reducing hospital readmission rates.

We aim to assess the causal effect of being discharged to an acute rehabilitation center versus home with home health services on hospital readmission rates through a retrospective observational study. Hospital records for acute medical and surgical patients discharged from three hospitals in the University of Pennsylvania Hospital system between 2010 and 2012 were collected; see Jones et al. (2015) for more details on this study. Within this data set, there are 4893 individuals assigned to acute rehabilitation and 35,174 individuals assigned to home with home health services, for 40,067 total individuals. We would like to assess whether discharge to acute rehabilitation reduces the causal risk of hospital readmission relative to discharge home with home health services. Beyond testing this hypothesis, we

would also like to create confidence intervals for causal parameters that effectively summarize the impact of discharge location on hospital readmission rates in our study population. Two causal estimands of interest for this comparison are the *causal risk difference*, which is the difference in proportions of readmitted patients if all patients had been assigned to acute rehabilitation versus that if all patients had been discharged home with home health services; and the *causal risk ratio*, which is the ratio of these two proportions.

Through the use of matching with a variable number of controls (Ming and Rosenbaum, 2000), individuals assigned to acute rehabilitation were placed in matched sets with varying numbers of home with home health services individuals (ranging from 1 to 20) who were similar on the basis of their observed covariates. We used rank-based Mahalanobis distance with a propensity score caliper (estimated by logistic regression) of 0.2 as our distance metric to perform the matching. We further required exact balance on the indicator of admission to an intensive care unit to better control for whether an individual had a critical illness. In Appendix B.1, we demonstrate that this stratification resulted in acceptable balance on the basis of the standardized differences between the groups.

In the stratified experiment that our match aims to mimic, randomization inference can be readily used to test Fisher's sharp null of no effect. Under Fisher's sharp null, the unobserved potential outcomes are assumed to equal the observed potential outcomes for each individual. The sharp null can then be assessed by noting that within each stratum, the number of treated individuals for whom an event is observed follows a hypergeometric distribution. The total number of treated individuals with events across all strata is then distributed as the sum of independent hypergeometric distributions, forming the basis for what has become known as the Mantel-Haenszel test (Mantel and Haenszel, 1959; Rosenbaum, 2002a).

Testing a null on the causal risk difference or the causal risk ratio presents challenges not encountered when testing the sharp null, as many allocations of potential outcomes could yield the same causal parameter. For example, if we are testing the null that the causal risk difference is 0 without making further assumptions on the potential outcomes, the

allocation under Fisher's null is merely one of many choices (i.e., it is merely one element of the composite null). Conducting a hypothesis test and performing a sensitivity analysis requires assessing tail probabilities for all elements of the composite null, both under the assumption of no unmeasured confounding and while allowing for an unmeasured confounder of a range of strengths. Direct enumeration of all possible combinations of potential outcomes is computationally infeasible for even moderate sample sizes. In our motivating example, there are $2^{40,067}$ possible combinations of potential outcomes, even *without* considering values for the unmeasured confounder.

We instead aim to find the combination of potential outcomes and unmeasured confounders that results in the worst-case p-value for the test being conducted. If the null hypothesis corresponding to this worst-case allocation can be rejected, we can then reject all elements of the composite null. Rosenbaum (2002b) uses a similar approach for inference on the *attributable effect*, a quantity which is closely related to the risk difference. There it is shown that under the assumption of a nonnegative treatment effect (i.e., the treatment may cause an event, but does not preclude an event from happening if it would have happened under the control) a simple enumerative algorithm yields an asymptotic approximation to the worst-case p-value for this composite null. This is because the impact on the p-value of attributing an observed event to the treatment (stating that the unobserved potential outcome under control is 0) can be well approximated through asymptotic separability (Gastwirth et al., 2000), such that one can satisfy the null while finding the worst-case allocation by sorting the strata on the basis of their impact on the p-value and attributing the proper number of effects by proceeding down the sorted list. Recent works by Yang et al. (2014) and Keele et al. (2014) discuss how the attributable effect can also be used to define estimands of interest in instrumental variable studies.

Unfortunately, in the absence of a known direction of effect finding the worst-case allocation does not simplify in the same manner. This is because finding the potential outcome allocation with the largest impact on the p-value on a stratum-wise basis does not readily yield an

allocation that satisfies the composite null. The problem is not separable on a stratum-wise basic even asymptotically, as the requirement that the composite null must be true necessarily links the strata together in a complex manner. There are two non-complementary forces at play in the required optimization problem: for some strata, the potential outcome allocations should maximize the impact on the p-value, while in other strata the missing potential outcome allocations should work towards satisfying the composite null. For our motivating example, there are over 300,000 types of contributions to the p-value that must be considered in the sensitivity analysis when we do not assume a known direction of effect (as is shown in Section 3.6.1). Explicit enumeration is intractable here, as we must consider which allowed *combinations* of these contributions maximize the p-value while satisfying the null in question. As such, a different approach is required to make the computation feasible.

3.1.3. Integer Programming as a Path Forward

In this paper, we show that hypothesis testing for a composite null with binary outcomes can be performed by solving an integer linear program under the assumption of no unmeasured confounding. When conducting a sensitivity analysis by allowing for unmeasured confounding of a certain strength, an integer quadratic program is required. These optimization problems yield the worst-case p-value within the composite null so long as a normal approximation to the test statistic is justified. We show that our formulation is strong, in that the optimal objective value for our integer program closely approximates that of the corresponding continuous relaxation. As we demonstrate through simulation studies and real data examples, this allows hypothesis testing and sensitivity analyses to be conducted efficiently even with large sample sizes despite the fact that integer programming is $\mathcal{NP}$-hard in general, as discrete optimization solvers heavily utilize continuous relaxations in their search path. Through comparing our formulation to an equivalent binary program in the supplementary material, we also demonstrate that recent advances in optimization software (Jünger et al., 2009) alone are not sufficient for solving the problem presented herein; rather, a thoughtful formulation remains essential for solving large-scale discrete

optimization problems expeditiously.

3.2. Causal Inference after Matching

3.2.1. Notation for a Stratified Randomized Experiment

Suppose there are I independent strata, the i^{th} of which contains $n_i \geq 2$ individuals, that were formed on the basis of pre-treatment covariates. In each stratum, m_i individuals receive the treatment and $n_i - m_i$ individuals receive the control, and $\min\{m_i, n_i - m_i\} = 1$. We proceed under the stable unit treatment value assumption (SUTVA), which entails that (1) there is no interference, i.e. that the observation of one unit is not affected by the treatment assignment of other units; and (2) there are no hidden levels of the assigned treatment, meaning that the treatments for all individuals with the same level of observed treatment are truly comparable (Rubin, 1986). Let Z_{ij} be an indicator variable that takes the value 1 if individual j in stratum i is assigned to the treatment. Each individual has two sets of binary potential outcomes: one under treatment, $\{r_{Tij}, d_{Tij}\}$, and one under control, $\{r_{Cij}, d_{Cij}\}$. r_{Tij} and r_{Cij} are the primary outcomes of interest, while d_{Tij} and d_{Cij} are indicators of whether or not an individual would actually take the treatment when randomly assigned to the treatment or control group. The observations for each individual are $R_{ij} = r_{Tij}Z_{ij} + r_{Cij}(1 - Z_{ij})$ and $D_{ij} = d_{Tij}Z_{ij} + d_{Cij}(1 - Z_{ij})$; see Neyman (1923) and Rubin (1974) for more on the potential outcomes framework. In the classical experimental setting, $d_{Tij} - d_{Cij} = 1 \; \forall i, j$, and hence all individuals take the administered treatment. For a randomized encouragement design, Z_{ij} represents the encouragement to take the treatment (which is randomly assigned to patients), while d_{Tij} and d_{Cij} are the actual treatment received if $Z_{ij} = 1$ and $Z_{ij} = 0$ respectively (Holland, 1988). Matched observational studies assuming strong ignorability (Rosenbaum and Rubin, 1983) aim to replicate a classical stratified experiment, whereas matched studies employing an instrumental variable strive towards a randomized encouragement design, with Z_{ij} being the instrumental variable.

There are $N = \sum_{i=1}^{I} n_i$ total individuals in the study. Each individual has observed covari-

ates $\mathbf{x}_{ij}$ and unobserved covariate u_{ij}. Let $\mathbf{R} = [R_{11}, R_{12}, ..., R_{In_I}]^T$, $\mathbf{R}_i = [R_{i1}, ..., R_{in_i}]^T$, and let the analogous definitions hold for $\mathbf{D}, \mathbf{D}_i$, $\mathbf{Z}, \mathbf{Z}_i$. Let $\mathbf{r}_T = [r_{T11}, ..., r_{TIn_I}]$, $\mathbf{r}_{Ti} = [r_{Ti1}, ..., r_{Tin_i}]$, and let the analogous definitions hold for the other potential outcomes and the unobserved covariate. Let $\mathbf{X}$ be a matrix whose rows are the vectors $\mathbf{x}_{ij}$. Finally, let Ω be the set of $\prod_{i=1}^{I} n_i$ possible values of $\mathbf{Z}$ under the given stratification. In a randomized experiment, randomness is modeled through the assignment vector; each $\mathbf{z} \in \Omega$ has probability $1/|\Omega|$ of being selected, where the notation $|B|$ denotes the number of elements in the set B. Hence, quantities dependent on the assignment vector such as $\mathbf{Z}$, $\mathbf{R}$ and $\mathbf{D}$ are random, whereas $\mathcal{F} = \{\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C, \mathbf{X}, \mathbf{u}\}$ contains fixed quantities. For a randomized experiment, $\mathbb{P}(Z_{ij} = 1|\mathcal{F}, \mathbf{Z} \in \Omega) = m_i/n_i$, and $\mathbb{P}(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathbf{Z} \in \Omega) = 1/|\Omega|$.

3.2.2. Conducting a Sensitivity Analysis

In an observational study, the I strata are still generated based on pre-treatment covariates but are only created *after* treatment assignment has taken place. Furthermore, the treatment assignment was conducted outside of the practitioner's control which may introduce bias due to the existence of unmeasured confounders. We follow the model for a sensitivity analysis of Rosenbaum (2002a, Section 4), which states that failure to account for unobserved covariates may result in biased treatment assignments within a stratum. This model can be parameterized by a number $\Gamma = \exp(\gamma) \geq 1$ which bounds the extent to which the odds ratio of assignment can vary between two individuals in the same matched stratum. Letting $\pi_{ij} = \mathbb{P}(Z_{ij} = 1|\mathcal{F})$, we can write the allowed deviation as $1/\Gamma \leq \pi_{ij}(1-\pi_{ik})/(\pi_{ik}(1-\pi_{ij})) \leq \Gamma$. This model can be equivalently expressed in terms of the observed covariates $\mathbf{x}_{ij}$ and the unobserved covariate u_{ij} (assumed without loss of generality to be between 0 and 1), as $\log(\pi_{ij}/(1-\pi_{ij})) = \zeta(\mathbf{x}_{ij}) + \gamma u_{ij}$, where $\zeta(\mathbf{x}_{ij}) = \zeta(\mathbf{x}_{ik}), i = 1, ..., I, 1 \leq j, k \leq n_i$. See Rosenbaum (2002a, Section 4.2) for a discussion of the equivalence between these models. The probabilities of each possible allocation of treatment and control are given by $\mathbb{P}(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathbf{Z} \in \Omega) = \exp(\gamma \mathbf{z}^T \mathbf{u}) / \sum_{\mathbf{b} \in \Omega} \exp(\gamma \mathbf{b}^T \mathbf{u})$, where $\mathbf{u} = [u_{11}, u_{12}, ..., u_{I,n_i}]$. If $\Gamma = 1$, the distribution of treatment assignments corresponds to the randomization distribution

discussed in Section 4.2.1. For $\Gamma > 1$, the resulting distribution differs from that of a randomized experiment with the extent of the departure controlled by Γ.

Consider a simple hypothesis test based on a test statistic of the form $T = \mathbf{Z}^T\mathbf{q}$, where $\mathbf{q} = \mathbf{q}(\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C)$ is a permutation invariant, arrangement increasing function. Most commonly employed statistics are of this form; see Rosenbaum (2002a, Section 2.4) for a detailed discussion. Without loss of generality reorder the elements of $\mathbf{q}$ such that within each stratum $q_{i1} \leq q_{i2} \leq .. \leq q_{in_i}$. For a given value of Γ and for fixed values of the potential outcomes, a sensitivity analysis proceeds by finding tight upper and lower bounds on the upper tail probability, $\mathbb{P}(T \geq t)$, by finding the worst-case allocation of the unmeasured confounder $\mathbf{u}$. One then finds the value of Γ such that the conclusions of the study would be materially altered. The more robust a given study is to unmeasured confounding, the larger the value of Γ must be to alter its findings.

As is demonstrated in Rosenbaum and Krieger (1990) for strata with $m_i = 1$, for each Γ an upper bound on $\mathbb{P}(T \geq t)$ is found at a value of the unobserved covariate $\mathbf{u}^+ \in \mathbf{U}_1^+ \times ... \times \mathbf{U}_I^+$, where $\mathbf{U}_i^+$ consists of $n_i - 1$ ordered binary vectors (each of length n_i) with $0 = u_{i1}^+ \leq u_{i2}^+ ... \leq u_{in_i}^+ = 1$. Similarly, a lower bound on $\mathbb{P}(T \geq t)$ is found at a vector $\mathbf{u}^- \in \mathbf{U}_1^- \times ... \times \mathbf{U}_I^-$ with $1 = u_{i1}^- \geq u_{i2}^- ... \geq u_{in_i}^- = 0$. Under mild regularity conditions on $\mathbf{q}$, T is well approximated by a normal distribution. Large sample bounds on the tail probability can be expressed in terms of corresponding bounds on standardized deviates. These results can readily extended to stratifications yielded by a full match through a simple redefinition of $\mathbf{Z}$ and $\mathbf{q}$; see Rosenbaum (2002a, Section 4, Problem 12).

3.3. Composite Null Hypotheses

3.3.1. Estimands of Interest

To motivate our discussion, we will focus on three causal estimands of interest with binary outcomes. Note however that the general framework for inference and sensitivity analyses presented herein can be applied to any causal estimand for binary potential outcomes with

an associated test statistic that can be written as $\mathbf{Z}^T\mathbf{q}$ for a function $\mathbf{q}(\cdot)$ that satisfies the conditions outlined in Section 3.2.2. The causal parameters we will consider are the causal risk difference, causal risk ratio, and the effect ratio, defined as:

$$\begin{aligned}
\text{Risk Difference} \quad & \delta := \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i}(r_{Tij} - r_{Cij}) \\
\text{Risk Ratio} \quad & \varphi := \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} r_{Tij}}{\sum_{i=1}^{I}\sum_{j=1}^{n_i} r_{Cij}} \\
\text{Effect Ratio} \quad & \lambda := \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(r_{Tij} - r_{Cij})}{\sum_{i=1}^{I}\sum_{j=1}^{n_i}(d_{Tij} - d_{Cij})}.
\end{aligned}$$

As mentioned in the introduction, the causal risk difference measures the difference in proportions of observed events had all the individuals received the treatment and observed events had all individuals received the control. Similarly, the causal risk ratio measures the ratio of these two proportions. Each of these estimands has merits and shortcomings relative to the other, owing to the fact that the risk difference measures an effect on an absolute scale while the risk ratio measures an effect on a relative scale; see Appendix B.2 for further discussion of these two measures. These estimands are appropriate under strong ignorability (Rosenbaum and Rubin, 1983); in the corresponding idealized experiment, there are simply treated and control individuals, and all individuals comply with their assigned treatment regimen.

The effect ratio is a ratio of two average treatment effects, and hence serves as an assessment of the relative magnitude of the two treatment effects (Baiocchi et al., 2010; Yang et al., 2014). It is a causal estimand of interest in instrumental variable studies. In the idealized experiment being mimicked, Z_{ij} represents the randomized encouragement to take the treatment or control, while d_{Tij} and d_{Cij} indicate whether the treatment would be taken if $Z_{ij} = 1$ and $Z_{ij} = 0$ respectively. The effect ratio then represents the ratio of the effect of the encouragement on the outcome to the effect of the encouragement on the treatment received. If the encouragement (1) is truly randomly assigned within strata defined by the

observed covariates; and (2) can only impact the outcome of an individual if the encouragement changes the individual's choice of treatment regimen (the *exclusion restriction*: $d_{Tij} = d_{Cij} \Rightarrow r_{Tij} = r_{Cij}$), $\mathbf{Z}$ is then an instrument for the impact of the treatment on the response (Angrist et al., 1996). The parameter λ still has an interpretation in terms of relative magnitude of the two effects even if the exclusion restriction is not met, but the exclusion restriction coupled with monotonicity ($d_{Tij} \geq d_{Cij}$, also referred to as assuming "no defiers") give λ an additional interpretation as the average treatment effect among individuals who are *compliers*, i.e. individuals for which $d_{Tij} - d_{Cij}$; this is commonly referred to as the *local* average treatment effect. While we will not always assume monotonicty holds, we will make the assumption that the encouragement has an *aggregate* positive effect, i.e. $\sum_{i=1}^{I} \sum_{j=1}^{n_i} d_{Tij} - d_{Cij} > 0$, such that the effect ratio is well defined.

3.3.2. Testing a Composite Null

Note first that a null hypothesis on δ, φ, or λ corresponds to a composite null hypothesis on the values of the potential outcomes, as multiple potential outcome allocations yield the same value for the causal parameter. Let $\Theta(\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C)$ be a function that maps a given set of potential outcomes to the corresponding causal parameter value of interest, θ. We call a set of potential outcomes $\{\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C\}$ *consistent* with a null hypothesis $H_0 : \theta = \theta_0$ for a causal parameter θ if the following conditions are satisfied:

(A1) Consistency with observed data: $Z_{ij} r_{Tij} + (1 - Z_{ij}) r_{Cij} = R_{ij}$; $Z_{ij} d_{Tij} + (1 - Z_{ij}) d_{Cij} = D_{ij}$

(A2) Consistency with assumptions made on potential outcomes

(A3) Agreement with the null hypothesis: $\Theta(\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C) = \theta_0$

The first condition recognizes that we know the true values for half of the potential outcomes based on the observed data. The second condition means that if the practitioner has made additional assumptions on the potential outcomes, those assumptions must be satisfied in

the allocations of potential outcomes under consideration. Assumptions could include a known direction of effect, monotonicity, the exclusion restriction, and combinations thereof. The third condition signifies that when testing a null hypothesis, we must only consider allocations of potential outcomes where the corresponding causal parameter takes on the desired value.

Let $\mathcal{H}(\theta_0)$ represent the set of potential outcomes satisfying conditions A1 - A3. As the size of a composite null hypothesis test is the supremum of the sizes of the elements of the composite null, to reject the null $H_0 : \theta = \theta_0$ at level α, we must reject the null for all $\{\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C\} \in \mathcal{H}(\theta_0)$ at level α. As direct enumeration of $\mathcal{H}(\theta_0)$ is a laborious (and likely computationally infeasible) task, we instead aim to find a single worst-case allocation $\{\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C\}^*$ such that rejection of $\{\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C\}^*$ at level α implies rejection for all $\{\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C\} \in \mathcal{H}(\theta_0)$.

We consider test statistics of the form $T(\theta_0) = \sum_{i=1}^{I} T_i(\theta_0)$ with expectation 0 under the null at $\Gamma = 1$. Let $\psi(\theta_0; \mathbf{r}_{Ti}, \mathbf{r}_{Ci}, \mathbf{d}_{Ti}, \mathbf{d}_{Ci}) = \mathbb{E}[T_i(\theta_0)]$. Thus, $\sum_{i=1}^{I} \psi(\theta_0; \mathbf{r}_{Ti},\ \mathbf{r}_{Ci}, \mathbf{d}_{Ti}, \mathbf{d}_{Ci}) = 0$ if and only if $\Theta(\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C) = \theta_0$. For our three estimands of interest, the stratum-wise contributions to the test statistic are

$$
\begin{aligned}
T_i(\delta_0) &= -n_i\delta_0 + n_i \sum_{j=1}^{n_i} (Z_{ij}R_{ij}/m_i - (1 - Z_{ij})R_{ij}/(n_i - m_i)) \\
T_i(\varphi_0) &= n_i \sum_{j=1}^{n_i} (Z_{ij}R_{ij}/m_i - \varphi_0(1 - Z_{ij})R_{ij}/(n_i - m_i)) \\
T_i(\lambda_0) &= n_i \sum_{j=1}^{n_i} (Z_{ij}(R_{ij} - \lambda_0 D_{ij})/m_i - (1 - Z_{ij})(R_{ij} - \lambda_0 D_{ij})/(n_i - m_i)) \,,
\end{aligned}
$$

with respective stratum-wise expectations

$$\psi(\delta_0; \mathbf{r}_{Ti}, \mathbf{r}_{Ci}, \mathbf{d}_{Ti}, \mathbf{d}_{Ci}) = -n_i\delta_0 + \sum_{j=1}^{n_i}(r_{Tij} - r_{Cij})$$
$$\psi(\varphi_0; \mathbf{r}_{Ti}, \mathbf{r}_{Ci}, \mathbf{d}_{Ti}, \mathbf{d}_{Ci}) = \sum_{j=1}^{n_i}(r_{Tij} - \varphi_0 r_{Cij})$$
$$\psi(\lambda_0; \mathbf{r}_{Ti}, \mathbf{r}_{Ci}, \mathbf{d}_{Ti}, \mathbf{d}_{Ci}) = \sum_{j=1}^{n_i}(r_{Tij} - \lambda_0 d_{Tij} - (r_{Cij} - \lambda_0 d_{Cij})).$$

To express these statistics in the required form for conducting a sensitivity analysis, define $\tilde{\mathbf{Z}}$ such that $\tilde{Z}_{ij} = Z_{ij}$ if $m_i = 1$ and $\tilde{Z}_{ij} = 1 - Z_{ij}$ if $m_i > 1$. If $m_i = 1$, define $\mathbf{q}(\cdot)$ as:

$$(\mathbf{q}(\delta_0; \mathbf{r}_{Ti}, \mathbf{r}_{Ci}, \mathbf{d}_{Ti}, \mathbf{d}_{Ci}))_j = n_i \left(-\delta_0 + r_{Tij}/m_i - \sum_{k \neq j} r_{Cik}/(n_i - m_i) \right)$$
$$(\mathbf{q}(\varphi_0; \mathbf{r}_{Ti}, \mathbf{r}_{Ci}, \mathbf{d}_{Ti}, \mathbf{d}_{Ci}))_j = n_i \left(r_{Tij}/m_i - \sum_{k \neq j} \varphi_0 r_{Cik}/(n_i - m_i) \right)$$
$$(\mathbf{q}(\lambda_0; \mathbf{r}_{Ti}, \mathbf{r}_{Ci}, \mathbf{d}_{Ti}, \mathbf{d}_{Ci}))_j = n_i \left((r_{Tij} - \lambda_0 d_{Tij})/m_i - \sum_{k \neq j} (r_{Cik} - \lambda_0 d_{Cik})/(n_i - m_i) \right).$$

The analogous definition holds when $m_i > 1$: simply redefine $\mathbf{q}(\cdot)$ within stratum i such that the proper contribution is given to $T_i(\cdot)$ if unit j in stratum i receives the control (and thus, all other units receive the treatment). The test statistic $\tilde{\mathbf{Z}}^T\mathbf{q}(\cdot)$ then has the required form for conducting a sensitivity analysis.

Under mild regularity conditions, Lyapunov's central limit theorem yields that all three of the test statistics $T(\theta_0)$ under consideration are well approximated by a normal distribution for $\Gamma \geq 1$. See Chapter 2 for a discussion with regards to the risk difference (the risk ratio follows through similar arguments), and see Baiocchi et al. (2010) for a discussion for the effect ratio. Finding the worst-case allocation $\{\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C\}^*$ at a given Γ can be well approximated by finding the allocation of potential outcomes and unobserved confounder that results in the worst-case standardized deviate. While this observation simplifies our task, it alone is not sufficient for making both inference and sensitivity analyses feasible

for our estimands of interest; rather, we must exploit other features of the optimization problem.

3.4. Symmetric Tables

We now introduce the required framework and notation for our optimization problem. Though many equivalent formulations are possible, the one we describe has a decision variable for each unique distribution on a stratum's contribution to the test statistic. This is an extension of the formulation given in Chapter 2, which was catered towards maximizing the variance of the estimated causal risk difference under no unmeasured confounding. In Section 3.5.3, we discuss the elements of our formulation which facilitate solving the corresponding integer program efficiently.

Let $\mathcal{T}_i^{zrd} = \{j : Z_{ij} = z, R_{ij} = r, D_{ij} = d\}$, $(z, r, d) \in \{0, 1\}^3$, $i \in \{1, ..., I\}$, denote the eight possible partitions of indices of individuals in stratum i into sets based on their value of the encouraged treatment, observed response, and taken treatment. Within each set, all members share the same value of either r_{Tij} or r_{Cij}, and of either d_{Tij} or d_{Cij}. For example, if $j, k \in \mathcal{T}_i^{011}$, then $r_{Cij} = r_{Cik} = d_{Cij} = d_{Cik} = 1$, yet the values of $r_{Tij}, r_{Tik}, d_{Tij}, d_{Tik}$ are unknown. Note that for the stratifications under consideration $\sum_{(r,d)\in\{0,1\}^2} |\mathcal{T}_i^{0rd}| = n_i - m_i$, $\sum_{(r,d)\in\{0,1\}^2} |\mathcal{T}_i^{1rd}| = m_i$, and the minimum of these two quantities is always 1. $|\mathcal{T}_i^{zrd}|$ can be thought of as the value in cell (z, r, d) of a 2^3 factorial table that counts the number of individuals with each combination of (z, r, d) in stratum i.

Under no assumption on the structure of the potential outcomes, there are 2^{2n_i} possible sets of potential outcomes in stratum i that are consistent with the observed data, each of which results in a particular distribution for the contribution to the test statistic from stratum i, $T_i(\theta_0)$. Fortunately, one need never consider all 2^{2n_i} allocations. First, without any assumptions on the potential outcomes, the 2^{2n_i} possible sets of potential outcomes in stratum i only yield $\prod_{(z,r,d)\in\{0,1\}^3}(|\mathcal{T}_i^{zrd}| + 1)^2$ unique distributions for $T_i(\theta_0)$. To see this, note that the test statistics under consideration are permutation invariant within each

stratum. Let us examine the set $\mathcal{T}_i^{000}$ as an illustration. Here, we have $d_{Cij} = r_{Cij} = 0$ for all $j \in \mathcal{T}_i^{000}$. Of the $2^{|\mathcal{T}_i^{000}|}$ pairings $[r_{Tij}, r_{Cij}]$, there are only $|\mathcal{T}_i^{000}| + 1$ non-exchangeable allocations of values for $\{r_{Tij} : j \in \mathcal{T}_i^{000}\}$: $(0, 0..., 0), (1, 0, ..., 0), ...,$ and $(1, 1, ..., 1)$. An analogous argument shows that there are only $|\mathcal{T}_i^{000}| + 1$ non-exchangeable arrangements for d_{Tij}, thus resulting in $(|\mathcal{T}_i^{000}| + 1)^2$ total non-exchangeable allocations. The same logic yields a contribution of $(|\mathcal{T}_i^{zrd}| + 1)^2$ for each of the other seven partitions.

Additional structure is often imposed on the potential outcomes on top of consistency with the observed data. For example, in the classical experiment we have that $d_{Tij} - d_{Cij} = 1 \; \forall i, j$, meaning that all patients comply with their assigned treatment. Hence, the four partitions where $Z_i - D_i \neq 0$ are empty, and in the remaining partitions d_{Tij} and d_{Cij} are fixed at 1 and 0 respectively. This results in only $\prod_{(z,r)\in\{0,1\}^2}(|\mathcal{T}_i^{zrz}| + 1)$ allowed non-exchangeable allocations within stratum i; note the lack of a square in the expression. This is also shown in Rigdon and Hudgens (2015, Section 3). Other assumptions such as a known direction of effect, monotonicity, and the exclusion restriction can be seen to similarly reduce the set of allowed non-exchangeable allocations.

It would seem as though we must consider at most $\prod_{i=1}^{I} \prod_{(z,r,d)\in\{0,1\}^3}(|\mathcal{T}_i^{zrd}| + 1)^2$ different distributions for $T(\theta_0) = \sum_{i=1}^{I} T_i(\theta_0)$ in our optimization problem. Fortunately, note first that we assume independence between strata, and further note that we are using a normal approximation to conduct inference. Hence, both the expectations and variances *sum* between strata and we do not need to consider covariances between strata. Further, in the same way that there were a limited number of non-exchangeable allocations of potential outcomes in each stratum due to repetition, many observed 2^3 factorial tables in the data are repeated multiple times. For example, the matching with multiple controls described in Section 3.1 returned 4893 strata, of which only 234 were unique.

3.4.1. Expectation, Variance, and Null Deviation

We now introduce the requisite notation to exploit these facts to facilitate inference. Let $\mathcal{C}_i = (|\mathcal{T}_i^{000}|, ..., |\mathcal{T}_i^{111}|)$ be the observed counts of the 2^3 tables for stratum i. $\mathfrak{C} = \{\mathcal{C}_1, ..., \mathcal{C}_I\}$ is a (multi)set, where the number of unique elements equals the number of unique 2^3 tables observed in the data, which will typically be much less than its dimension. Let S be the number of unique tables, and let $s \in \{1, ..., S\}$ index the unique tables. Define $\mathcal{I}(i)$ to be a function returning the index of the unique table corresponding to the table observed in stratum i. Hence, $\mathcal{I}(i) = \mathcal{I}(\ell)$ if and only if $\mathcal{C}_i = \mathcal{C}_\ell$. Let $M_s = |\mathcal{I}^{-1}(s)|$ be the number of strata where unique table s was observed, and let $\tilde{n}_s = n_b$ for any $b \in \mathcal{I}^{-1}(s)$ be the number of observations in unique table s. Finally, let P_s be the number of allowed non-exchangeable potential outcomes for unique table s, and let $\{[\mathbf{r}_{T[sp]}, \mathbf{r}_{C[sp]}, \mathbf{d}_{T[sp]}, \mathbf{d}_{C[sp]}]\}, p \in \{1, ..., P_s\}$ be the set of allowed potential outcome allocations that are consistent with unique table s, where tablewise consistency refers to adherence to conditions A2 and A3 within table s.

Without loss of generality, we assume that the observed statistic, t_{θ_0}, is larger than its expectation under the null at $\Gamma = 1$, 0. In upper bounding the upper tail probability $P(T(\theta_0) \geq t_{\theta_0})$, we thus restrict our search to the set of unobserved confounders $\mathbf{u}^+ \in \mathbf{U}^+$ as discussed in Section 3.2.2. The analogous procedure would hold for $\mathbf{u}^- \in \mathbf{U}^-$ if $t_{\theta_0} < 0$.

For the s^{th} unique table, and the p^{th} set of allowed potential outcome allocations consistent within table s, $s \in \{1, ..., S\}$, $p \in \{1, .., P_s\}$, form
$q(\theta_0)_{[sp]j} = (\mathbf{q}(\theta_0; \mathbf{r}_{T[sp]}, \mathbf{r}_{C[sp]}, \mathbf{d}_{T[sp]}, \mathbf{d}_{C[sp]}))_j$. Reorder the $q(\theta_0)_{[sp]j}$ such that $q(\theta_0)_{[sp]1} \leq q(\theta_0)_{[sp]2} \leq .. \leq q(\theta_0)_{[sp]\tilde{n}_s}$. For a given value of $\Gamma \geq 1$, we define $\mu(\theta_0)_{[sp]a}$ and $\nu(\theta_0)_{[sp]a}$, $a \in \{1, ...\tilde{n}_s - 1\}$, as

$$\mu(\theta_0)_{[sp]a} = \frac{\sum_{j=1}^{a} q(\theta_0)_{[sp]j} + \Gamma \sum_{j=a+1}^{\tilde{n}_s} q(\theta_0)_{[sp]j}}{a + \Gamma(\tilde{n}_s - a)}, \tag{3.1}$$

and

$$\nu(\theta_0)_{[sp]a} = \frac{\sum_{j=1}^{a}(q(\theta_0)_{[sp]j})^2 + \Gamma\sum_{j=a+1}^{\tilde{n}_s}(q(\theta_0)_{[sp]j})^2}{a + \Gamma(\tilde{n}_s - a)} - (\mu(\theta_0)_{[sp]a})^2. \tag{3.2}$$

This notation is reminiscent of that of Gastwirth et al. (2000). The index a corresponds to the the vector of unmeasured confounders $\mathbf{u}^+$ with a zeroes followed by $\tilde{n}_s - a$ ones. $\mu(\theta_0)_{[sp]a}$ and $\nu(\theta_0)_{[sp]a}$ represent the expectation and variance of the contribution to the test statistic $T(\theta_0)$ from a matched set with observed table s, consistent set of potential outcomes p, and allocation of unmeasured confounders a. Let $\boldsymbol{\mu}_{\theta_0} = [\mu(\theta_0)_{[11]1}, ..., \mu(\theta_0)_{[SP_S],\tilde{n}_S-1}]$, and let $\boldsymbol{\nu}_{\theta_0} = [\nu(\theta_0)_{[11]1}, ..., \nu(\theta_0)_{[SP_S],\tilde{n}_S-1}]$. Finally, recalling the definition of $\psi(\cdot)$ from Section 3.3 as the expectation of the contribution to the test statistic $T(\theta_0)$ from stratum i, define $\psi(\theta_0)_{[sp]j} = (\psi(\theta_0; \mathbf{r}_{T[sp]}, \mathbf{r}_{C[sp]}, \mathbf{d}_{T[sp]}, \mathbf{d}_{C[sp]}))_j$, and define $\boldsymbol{\psi}_{\theta_0} = [\psi(\theta_0)_{[11]1}, ..., \psi(\theta_0)_{[SP_S],\tilde{n}_S-1}]$.

3.5. Inference and Sensitivity Analysis

Let $x_{[sp]a}$ be an integer variable denoting how many times the set of potential outcomes p that is consistent with unique table s with allocation of unmeasured confounders a is observed in the data, $s \in \{1, ..., S\}$, $p \in \{1, ..., P_s\}$, $a \in \{1, ..., \tilde{n}_s - 1\}$, and let $\mathbf{x} = [x_{[11]1}, .., x_{[SP_s],\tilde{n}_S-1}]$. For a given θ_0 being tested, $\mu(\theta_0)_{[sp]a}x_{[sp]a}$ and $\nu(\theta_0)_{[sp]a}x_{[sp]a}$ represent the contribution to the overall mean and variance of the test statistic if the p^{th} set of potential outcomes in unique table s with allocation of unmeasured confounders a is observed $x_{[sp]a}$ times, and $\boldsymbol{\mu}_{\theta_0}^T\mathbf{x}$ and $\boldsymbol{\nu}_{\theta_0}^T\mathbf{x}$ represent the overall expectation and variance across all unique tables, potential outcomes and unmeasured confounders. $\sum_{p=1}^{P_s}\sum_{a=1}^{\tilde{n}_s-1} x_{[sp]a}$ then represents how many times the s^{th} unique table was observed in the data, a number which we defined to be M_s. Hence, $\sum_{p=1}^{P_s}\sum_{a=1}^{\tilde{n}_s-1} x_{[sp]a} = M_s$.

Note that through our formulation we have restricted optimization to the set of observations that adhere to conditions A1 (consistency with the observed data) and A2 (consistency with any other assumptions made by the modeler on the potential outcomes) of Section 3.3.2. We enforce condition A3 (that the null must be true in the resulting allocation of potential

outcomes) through adding a linear constraint to our optimization problem: $\boldsymbol{\psi}_{\theta_0}^T \mathbf{x} = 0$. The following integer program facilitates hypothesis testing and confidence interval construction under no unmeasured confounding (Section 3.5.1), as well as a sensitivity analysis for any $\Gamma > 1$ (Section 3.5.2):

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \;\; & (t_{\theta_0} - \boldsymbol{\mu}_{\theta_0}^T \mathbf{x})^2 - \kappa(\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}) \\ \text{subject to} \;\; & \sum_{p=1}^{P_s} \sum_{a=1}^{\tilde{n}_s - 1} x_{[sp]a} = M_s \;\; \forall s \\ & \boldsymbol{\psi}_{\theta_0}^T \mathbf{x} = 0 \\ & x_{[sp]a} \in \mathbb{Z} \;\; \forall s, p, a \\ & x_{[sp]a} \geq 0 \;\; \forall s, p, a, \end{aligned} \tag{P1}$$

where $\mathbb{Z}$ are the integers and $\kappa > 0$ is a positive constant to be described. The above formulation is sufficient for tests on the risk difference and risk ratio. For the effect ratio, we can impose the constraint of an aggregate positive effect of the intervention, $\sum_{i=1}^{I} \sum_{j=1}^{n_i} d_{Tij} - d_{Cij} > 0$, through an additional linear inequality.

3.5.1. Hypothesis Testing and Confidence Intervals Under No Unmeasured Confounding

For conducting inference under pure randomization (that is, under $\Gamma = 1$), the value of $\boldsymbol{\mu}_{\theta_0}^T \mathbf{x}$ is fixed to the expectation of the test statistic under the null, 0. Hence, $(t_{\theta_0} - \boldsymbol{\mu}_{\theta_0}^T \mathbf{x})$ is constant as well, and (MV) reduces to an integer linear program. This program is equivalent to finding the largest variance over all feasible $\mathbf{x}$. Call the optimal vector $\mathbf{x}_{\theta_0}^*$, and call the corresponding maximal variance $\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}_{\theta_0}^*$. The worst-case deviate for testing $\theta = \theta_0$ can then be found by setting $z_{\theta_0} = t_{\theta_0} / \sqrt{\nu_{\theta_0}^T \mathbf{x}_{\theta_0}^*}$.

To form a $100 \times (1 - \alpha)\%$ confidence interval at $\Gamma = 1$, we simply invert a series of tests. Explicitly, we find upper and lower bounds, θ_u and θ_ℓ, such that $\theta_\ell =$ $\textbf{SOLVE} \left\{ \theta : t_\theta / \sqrt{\nu_\theta^T \mathbf{x}_\theta^*} = z_{1-\alpha/2} \right\}$ and $\theta_u = \textbf{SOLVE} \left\{ \theta : t_\theta / \sqrt{\nu_\theta^T \mathbf{x}_\theta^*} = z_{\alpha/2} \right\}$, where z_q is the q quantile of a standard normal distribution. These endpoints can be found through a

grid search over θ, or by using the bisection algorithm.

3.5.2. Sensitivity Analysis through Iterative Optimization

For $\Gamma > 1$, (MV) is instead an integer quadratic program. First, note that we reject the null with a two-sided alternative at size α if $(t_{\theta_0} - \boldsymbol{\mu}_{\theta_0}^T \mathbf{x})^2 / (\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}) \geq \chi^2_{1,1-\alpha}$ for all values of the potential outcomes that are consistent with the null being tested, where $\chi^2_{1,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2_1 distribution. Equivalently, we need only determine whether $(t_{\theta_0} - \boldsymbol{\mu}_{\theta_0}^T \mathbf{x})^2 - \chi^2_{1,1-\alpha}(\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}) \geq 0$ for all feasible $\mathbf{x}$. This can be done by minimizing (MV) with $\kappa = \chi^2_{1,1-\alpha}$ over all feasible $\mathbf{x}$, and checking whether or not the objective value at $\mathbf{x}^*_{\theta_0}$ is greater than zero.

One may also be interested in knowing the worst-case deviate itself (equivalently, the worst-case p-value), rather than simply knowing the result of the test. The optimal vector $\mathbf{x}^*_{\theta_0}$ for (MV) at $\kappa = \chi^2_{1,1-\alpha}$ need not result in the worst-case deviate; however, we now show that we can find the worst-case p-value through an iterative procedure based on (MV). To proceed, we find the value $\kappa = \kappa^*$ such that the minimal objective value of (MV) equals 0. As is proved in Dinkelbach (1967), such a value of κ^* exactly equals the minimal squared deviate. Interpreted statistically, the value κ^* is the maximal critical value for the squared deviate such that the null could be still be rejected, which is equivalent to the value of the deviate itself. Although finding this zero could be performed using a grid search, we instead solve for the optimal $\mathbf{x}^*_{\theta_0}$ through the following algorithm.

1. Start with an initial value $\kappa^{(0)}$.
2. In iteration $i \geq 1$, set $\kappa = \kappa^{(i-1)}$ in (MV).
3. Solve the resulting program, and set $\kappa^{(i)} = (t_{\theta_0} - (\boldsymbol{\mu}_{\theta_0}^T \mathbf{x}_{\theta_0}^{*(i)}))^2 / (\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}_{\theta_0}^{*(i)})$.
4. If $\kappa^{(i)} = \kappa^{(i-1)}$ terminate the algorithm: set $\mathbf{x}^*_{\theta_0} = \mathbf{x}^{*(i)}_{\theta_0}$, and set $\kappa^* = \kappa^{(i)}$.
5. Otherwise, return to step 2. Repeat until convergence.

Note that the sequence $\{\kappa^{(i)}\}$ is bounded below by 0. It is also monotone decreasing for $i \geq 1$, as $(t_{\theta_0} - \boldsymbol{\mu}_{\theta_0}^T \mathbf{x}_{\theta_0}^{*(i+1)})^2 - \kappa^{(i)}(\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}_{\theta_0}^{*(i+1)}) \leq (t_{\theta_0} - \boldsymbol{\mu}_{\theta_0}^T \mathbf{x}_{\theta_0}^{*(i)})^2 - \kappa^{(i)}(\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}_{\theta_0}^{*(i)}) = 0$, which implies $\kappa^{(i)} \geq (t_{\theta_0} - (\boldsymbol{\mu}_{\theta_0}^T \mathbf{x}^{*(i+1)}))^2 / (\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}_{\theta_0}^{*(i+1)}) = \kappa^{(i+1)}$. Hence, this algorithm will converge to a stationary point κ^*. In practice, we find that this is achieved very quickly, frequently within 2 or 3 steps. At κ^*, note that it must be the case that the objective value in (MV) equals 0. This means that at the termination of the iterative procedure, we have converged to the minimal deviate. The maximal p-value is then $\Phi(-\sqrt{\kappa^*})$ for a one-sided test or $2 \times \Phi(-\sqrt{\kappa^*})$ for a two-sided test, where $\Phi(\cdot)$ is the CDF of a standard normal distribution.

3.5.3. Computation Time

In the past, researchers have been dissuaded from suggesting methodology that requires the solution of an integer program, as problems of this sort are $\mathcal{NP}$-hard in general. In this section, we present simulation studies to assuage fears that our integer linear ($\Gamma = 1$) and quadratic ($\Gamma > 1$) programs may have excessive computational burden. Before doing so, we discuss two properties of an integer programming formulation that substantially influence the performance of integer programming solvers: the strength of the corresponding continuous relaxation, and the avoidance of symmetric feasible solutions (Bertsimas and Tsitsiklis, 1997).

A strong formulation of an integer program is one for which the polyhedron defined by the constraint set, $\mathcal{P} = \{\mathbf{x} : \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \mathbb{R}\}$, is close to the integer hull, $\mathcal{P}_I = Conv\{\mathbf{x} : \mathbf{x} \in \mathcal{P} \cap \mathbb{Z}\}$. In an ideal world, the integer hull and the relaxed polyhedron would align, meaning that any *linear* programming relaxation would be guaranteed to have an integral optimal solution since any linear program has an optimal solution at the vertex of its corresponding polyhedron. For a quadratic program, having $\mathcal{P}_I = \mathcal{P}$ does not guarantee coincidence of the true and relaxed optimal solutions, as a quadratic program may have a solution at an edge. Nonetheless, having $\mathcal{P}$ far from $\mathcal{P}_I$ can hamper the progress of a mixed integer programming solver, as it increases the number of cuts required by branch-and-cut algorithms to strengthen the continuous relaxation (Mitchell, 2002).

A symmetric formulation is one in which variables can be permuted without changing the structure of the problem. Formulations of this sort can also cripple standard integer programming solvers even with modest problem size. This is due in large part to the generation of isomorphic solution paths by branch-and-bound and branch-and-cut algorithms, which in turn complicates the process by which a given node is proven optimal or suboptimal. Although methods exist to detect symmetry groups in a given formulation, formulations that explicitly avoid such groups are strongly preferred; see Margot (2010) for a discussion of these points.

We now present simulation studies to demonstrate that neither weakness nor symmetry of formulation proves inimical to conducting hypothesis testing and sensitivity analyses using the methodology outlined in this paper, even with large data sets and large stratum sizes. In our first setting, in each of 1000 iterations we sample 1250 matched sets from the strata in our motivating example from Section 3.1.2. We assign treated individuals and control indivduals an outcome of 1 with probability 0.75 and 0.25 respectively. Each iteration thus has strata ranging in size from 2 to 21, and each data set has an average of roughly 10,000 individuals within it. Large strata affect computation time, as they result in larger numbers of non-exchangeable potential outcome allocations within a stratum and fewer duplicated 2×2 tables in the data. In our data set, 25% of the matched strata had one acute rehabilitation individual and 20 home with home health services patients. This simulation setting thus produces particularly challenging optimization problems: on average, each iteration had 170,000 variables over which to optimize. As we demonstrate in Appendix B.3, the number of variables, itself affected by the number and size of the unique observed tables, is a primary determinant of computation time for the optimization routine.

We conduct two hypothesis tests in each iteration: a null on the causal risk difference, $\delta = 0.2$, and on the causal risk ratio, $\varphi = 1.75$. For both of the causal estimands being assessed, we test the stated nulls with two-sided alternatives at $\Gamma = 1$ (no unmeasured confounders, integer linear program) and $\Gamma = 3$ (unmeasured confounding exists, integer

quadratic program). We record the required computation time for each data set, which includes both the time taken to define the necessary constants for the problem and also the time required to solve the optimization problem. To measure the strength of our formulation, we also recorded whether or not the initial continuous relaxation had an optimal solution which was itself integral, and if not the relative difference in optimal objective function values between the integer and continuous formulations (defined to be the absolute difference of the two, divided by the absolute value of the relaxed value). Simulations were conducted on a desktop computer with a 3.40 GHz processor and 16.0 GB RAM. The `R` programming language was used to formulate the optimization problem, and the `R` interface to the `Gurobi` optimization suite was used to solve the optimization problem.

Table 12 shows the results of this simulation study. As one can see, our formulation yields optimal solutions in well under a minute for both the integer linear and integer quadratic formulations despite the magnitude of the problem at hand. The strength of our formulation is further evidenced by the typical discrepancy between the integer optimal solution and that of the continuous relaxation. For testing the causal risk difference, we found that in all of the simulations performed assuming no unmeasured confounding the integer program and its linear relaxation had the *same* optimal objective value. When testing at $\Gamma = 3$ the quadratic relaxation differed from the integer programming solution in roughly $2/3$ of the simulations; however, the resulting average relative gap between the two was a minuscule $3 \times 10^{-4}\%$. For testing the causal risk ratio, the objective values tended not to be identically equal at $\Gamma = 1$ or $\Gamma = 3$, which has to do with the existence of fractional values in the row of the constraint matrix enforcing the null hypothesis; nonetheless, the average gap among those iterations where there was a difference was $4 \times 10^{-5}\%$ for the linear program, and 0.002% for the quadratic program. This suggests not only that we have arrived upon a strong formulation, but that one could in practice accurately approximate (MV) by its continuous relaxation.

Appendix B.3 contains additional simulation studies which serve not only to further illustrate the strength of our formulation, but also to provide insight into what elements of the problem

Table 3: Computation times for tests of $\delta = 0.2$ and $\varphi = 1.75$ at $\Gamma = 1$ (integer linear program) and $\Gamma = 3$ (integer quadratic program), along with percentages of coincidence of the integer and relaxed objective values, and average gaps between integer solution and the continuous relaxation if a difference existed between the two.

H_0 Γ	Avg. Time (s), Integer	Avg. Time (s), Relaxation	$\%(obj_{int} = obj_{rel})$	Avg Gap If Different
$\delta = 0.2; \Gamma = 1$	5.88	5.59	100%	NA
$\delta = 0.2; \Gamma = 3$	9.77	7.14	36.9%	3×10^{-4}%
$\varphi = 1.75; \Gamma = 1$	5.86	5.62	0%	4×10^{-5}%
$\varphi = 1.75; \Gamma = 3$	10.85	7.82	3.2%	0.002%

affect computation time. We present simulations varying the value of Γ used, the number of matched sets, the null hypothesis being tested, the magnitude of the true effect, and the prevalence of the outcome under treatment and control in order to assess the impact of each of these factors on the time required to define the required constants and to carry out the optimization. We then compare our formulation to an equivalent, but highly symmetric, formulation in order to highlight the importance of avoiding symmetry for achieving a strong formulation with reasonable computation time. We also present a simulation study akin to the one presented in this section but using real data for the outcome variables as opposed to simulated outcomes. Finally, we provide advice for using our procedure under time constraints for the optimization routine.

3.6. Data Examples

We employ our methodology in two data examples. In Section 3.6.1, we present hypothesis testing and a sensitivity analysis for the causal risk difference and causal risk ratio in our motivating example from Section 3.1, wherein we compare hospital readmission rates for two different post-hospitalization protocols after an acute care hospitalization. In Section 3.6.2, we reexamine the instrumental variable study of Yang et al. (2014) comparing mortality rates for premature babies being delivered by *c*-section versus vaginal births. In addition to inference, confidence intervals, and sensitivity analyses, we also provide point estimators for the causal estimands of interest. These are formed by using our test statistic, $T(\theta)$, as an

estimating equation for an m-estimator (Van der Vaart, 2000), i.e $\hat{\theta} := \mathbf{SOLVE}\{\theta : T(\theta) = 0\}$; see Appendix B.4 for further discussion.

As will be shown, the findings in both of our examples exhibit varying degrees of sensitivity to unmeasured confounding: under the strongest assumptions, we fail to reject the null of no treatment effect after $\Gamma = 1.157$ in our first example and after $\Gamma = 1.67$ in our second. To provide context for the levels of robustness possible in a well designed observational study, Section 4.3.2 of Rosenbaum (2002a) notes that the finding of a causal relationship between smoking and lung cancer in Hammond (1964) continued to be significant until $\Gamma = 6$, meaning that an unmeasured confounder would have had to increase the odds of smoking by a factor of six while nearly perfectly predicting lung cancer in order to overturn the study's finding.

3.6.1. Risk Difference and Risk Ratio

We now return to our study of the impact of discharge to an acute rehabilitation center versus to home with home health services on hospital readmission rates after an acute care hospitalization. We use sixty day hospital readmission after initial hospital discharge as our outcome of interest. In terms of counterfactuals, we want to compare sixty day hospital readmission rates if all patients had been sent to acute rehabilitation with readmission rates if all patients had been assigned to home with home health services. We define $R_{ij} = 1$ if an individual was readmitted to the hospital, and 0 otherwise. We let $Z_{ij} = 1$ if an individual was assigned to acute rehabilitation. The marginal proportions of sixty day hospital readmission after accounting for observed confounders through matching are 0.206 for acute rehabilitation, and 0.243 for home with home health services. We will analyze this data set with and without the assumption of a known direction of effect. When assuming a direction of effect we assume that it is nonpositive in this example, meaning that going to acute rehabilitation can never hurt an individual: an individual who would not be readmitted to the hospital within sixty days after being discharged to home with home health services could not have been readmitted to the hospital within sixty days after being discharged to

acute rehabilitation.

The estimated risk difference is $\hat{\delta} = -0.0369$ (favoring acute rehabilitation) regardless of whether we assume a nonpositive treatment effect. We construct confidence intervals by inverting a series of hypothesis tests on $\{\delta_0\}$. Without assuming a nonpositive treatment effect, we find a 95% confidence interval for δ of [-0.0557; -0.0175]. With the assumption of a nonpositive effect, the 95% confidence interval shrinks to [-0.0535; -0.0202]. We conduct inference on the risk ratio, φ, in a similar manner. The estimated risk ratio was $\hat{\varphi} = 0.848$ (favoring acute rehabilitation); 95% confidence intervals for φ are [0.773; 0.927] and [0.780; 0.916] without and with assuming a nonpositive treatment effect respectively.

The results of a sensitivity analysis for a test of $\delta = 0 \Leftrightarrow \varphi = 1$ with a lower one-sided alternative are shown in Table 4. As one can see, the result is sensitive to unobserved biases under both scenarios, but far more so when we do not make an assumption on the direction of effect. To better understand this, it is useful to think of the corresponding integer programs that result in these worst-case bounds. The optimization problem with the assumption of a nonpositive treatment effect has 2,830 variables associated with it, with variables only corresponding to a choice of vector $\mathbf{u}_i^-$ in a given stratum. Without making this assumption, the number of variables grows to 321,860, as we must consider all non-exchangeable allocations of potential outcomes *and* all choices for the vector of unmeasured confounders. The difference in problem size impacts not only robustness against unmeasured confounding, but also computation time. The computations for each value of $\Gamma > 1$ shown took an average of 1.5 seconds under the assumption of non-negativity, but 75 seconds without this assumption. See Appendix B.5 for a discussion of why the assumption of a known direction of effect has such a substantial impact. Considering the sheer size of the problem, this bears testament to the strength of our formulation: for all of the Γ values tested, the continuous relaxation had an integer solution.

Table 4: Sensitivity analysis for a one-sided test with alternative hypothesis $\delta < 0 \Leftrightarrow \varphi < 1$. Worst case p-values are shown with (rightmost column) and without (middle column) assuming a known direction of effect.

Γ	$r_{Tij} \gtreqless r_{Cij}$	$r_{Tij} \leq r_{Cij}$
1.000	1.0×10^{-4}	6.1×10^{-6}
1.080	0.0306	0.0016
1.095	0.050	0.0028
1.157	0.420	0.050

3.6.2. Effect Ratio

Yang et al. (2014) present an observational study comparing the effect of cesarian section versus vaginal delivery on the survival of premature babies of 23-24 weeks gestational age, where $R_{ij} = 1$ if a baby survives. The analysis used whether or not a baby was delivered at a hospital with "high" rates of c-section as a potential instrumental variable. We present a sensitivity analysis for these data under combinations of assumptions of varying strength. In so doing, we aim to assess the impact of various assumptions on the inference's perceived sensitivity to unmeasured confounding. 1489 pairs of babies were formed, with a baby in the "high" group being matched to baby in the "low" group who was similar on the basis of all other pre-treatment covariates. Let $Z_{ij} = 1$ if the baby was delivered at a hospital with a high c-section rate, and let $D_{ij} = 1$ if the baby was delivered by a c-section. As such, the "randomized encouragement" is the type of hospital at which the baby was delivered, and the treatment of interest is the actual method of delivery.

We present inference on the effect ratio under all eight combinations of enforcing and not enforcing a nonnegative direction of effect (DE) : $r_{Tij} \geq r_{Cij}\ \forall i, j$; monotonicty (MO): $d_{Tij} \geq d_{Cij}\ \forall i, j$, and the exclusion restriction (ER): $d_{Tij} = d_{Cij} \Rightarrow r_{Tij} = r_{Cij}\ \forall i, j$. In the context of this example, the effect ratio is the ratio of the increase in survival rate to the increase in rate of c-sections for premature babies of 23-24 weeks gestational age that occurs with being delivered at a hospital with a high rate of c-sections. If we additionally assume that both monotonicity and the exclusion restriction hold, then the effect ratio has

Table 5: Minimal value of Γ such that conclusion of the hypothesis test on λ is reversed under eight combinations of assumptions.

$H_0 : \lambda = 0$	No (DE)	No (DE)	Yes (DE)	Yes (DE)
	No (MO)	Yes (MO)	No (MO)	Yes (MO)
No (ER)	1.292	1.292	1.677	1.677
Yes (ER)	1.292	1.371	1.677	1.677
$H_0 : \lambda = 0.1$	No (DE)	No (DE)	Yes (DE)	Yes (DE)
	No (MO)	Yes (MO)	No (MO)	Yes (MO)
No (ER)	1.213	1.220	1.407	1.409
Yes (ER)	1.225	1.270	1.408	1.410

the additional interpretation of being the effect of delivering at a hospital with high rates of c-sections among babies who would have been delivered by c-section if and only if they were delivered at a hospital with a high rate of c-sections.

Under any combination of assumptions, the estimated effect ratio is $\hat{\lambda} = 0.866$. Assuming none of (DE), (MO), (ER), the 95% confidence interval is [0.50; 1.47], and there are 256 decision variables in the optimization problem. Assuming all of (DE), (MO), (ER), the 95% confidence interval shrinks to [0.58; 1], and there are 49 decision variables in the optimization problem.

In Table 6, we present the values of Γ required to overturn the rejection of the nulls that $\lambda = 0$ and $\lambda = 0.1$, both with an upper one-sided alternative at $\alpha = 0.05$. For the null of $\lambda = 0$, this test boils down to a test on the average treatment effect, but with a range of restrictions on the potential outcomes. Once a nonnegative direction of effect is imposed (the bottom four cells of the table), the test of $\lambda = 0$ simply becomes a test of Fisher's sharp null; see Appendix B.5 for further discussion. Because of this, the assumptions of monotonicity and the exclusion restriction cannot impact the sensitivity analysis at $\lambda = 0$ unless non-negativity is not enforced. Furthermore, without assuming a direction of effect, monotonicity can only affect the performed inference if it is enforced in concert with the exclusion restriction at $\lambda = 0$ and vice versa. For $\lambda = 0.1$, the test no longer corresponds exclusively to one of Fisher's sharp null when non-negativity is imposed. We thus see that

each assumption impacts the study's robustness against unmeasured confounding to varying degrees. For all combinations of assumptions and each value of Γ tested, the corresponding integer quadratic program solved in under 2 seconds.

3.7. Discussion

Our formulation exploits attributes of the randomization distributions for our proposed test statistics which are unique to inference after matching. While this is sufficient for our purposes, one resulting limitation is that our method will likely not be practicable in observational studies or randomized clinical trials where there either are no strata, or where each stratum contain a large number of both treated *and* control individuals; see Rigdon and Hudgens (2015) for a discussion of the difficulties of conducting randomization inference with binary outcomes in these settings. In these settings, the work of Cornfield et al. (1959) presents a method for sensitivity analysis for the risk ratio, and Ding and Vanderweele (2014) extend this approach to the risk difference. Another limitation is that as with any $\mathcal{NP}$-hard endeavor, it is difficult to anticipate ahead of time how long our method will take on a given data set with a given match structure; however, through a host of simulation studies presented both in Section 3.5.3 and Appendix B.3 we have provided further insight into these matters for practitioners interested in using our methods.

We have framed hypothesis testing and sensitivity analyses for composite null hypotheses with binary outcomes in matched observational studies as the solutions to integer linear ($\Gamma = 1$) and quadratic ($\Gamma > 1$) programs. An interesting consequence of our formulation is that it readily yields a method for performing a sensitivity analysis for simple null hypotheses under general outcomes without reliance on the asymptotically separable algorithm of Gastwirth et al. (2000); see Appendix B.6 for details and a data example. We have shown that our method can be practicable even with large data sets and large stratum sizes. We have further demonstrated through simulation studies and real data examples that our formulation explicitly avoids issues known to hinder the performance of integer programming algorithms such as looseness of formulation and symmetry. In so doing, we hope to shed

further light on the usefulness of integer programming for solving problems in causal inference.

CHAPTER 4 : Sensitivity Analysis for Multiple Comparisons in Matched Observational Studies through Quadratically Constrained Linear Programming

Joint work with Dylan Small

4.1. Introduction

4.1.1. Unmeasured Confounding with Multiple Outcomes

Conclusions drawn from an observational study should be subjected to additional scrutiny due to their vulnerability to unmeasured confounding. Unlike with a randomized experiment, a covariate which has not been adjusted for in the primary analysis may very well drive the observed relationship, thus nullifying the study's original finding. This necessitates an additional step known as a *sensitivity analysis* to assess the robustness of an observational study's conclusions. A sensitivity analysis seeks an answer to the following question: how extreme would hidden bias have to be in order for the conclusions of a study to be materially altered? A study whose findings could be overturned with a small amount of unmeasured confounding invites warranted skepticism, while a study's conclusions are bolstered if a large degree of unmeasured confounding is required.

A sensitivity analysis computes worst-case bounds on the desired inference at a given level of unmeasured confounding. In observational studies employing matching to adjust for overt biases, the corresponding sensitivity analysis has been well studied when there is a single outcome variable of interest; see Section 4 of Rosenbaum (2002a) for a comprehensive overview. It is parameterized by a number $\Gamma \geq 1$ which controls the allowable departure from purely random assignment for individuals who are similar on the basis of their observed covariates: two individuals in the same matched set can, due to the presence of unmeasured confounding, differ in their odds of assignment to treatment by at most Γ. Higher values of Γ thus allow for unmeasured confounding to more substantially bias the treatment assignment

probabilities for individuals in the same matched set. As discussed in Section 4.2.2, the impact of unmeasured confounding can be encoded by a scalar latent variable, u_{ij}, which represents the aggregate impact of unmeasured confounding on the assignment probabilities for individual j in matched set i. Individuals in the same matched set with higher values for u_{ij} have higher probabilities of assignment to treatment. At each level of Γ, one finds the vector of unmeasured confounders for all individuals in the study, $\mathbf{u}$, which maximizes the p-value, hence yielding the worst possible inference for a given departure from purely random assignment.

Matched observational studies may seek to investigate the impact of a single treatment on multiple outcome variables; see Sabia (2006), Voigtländer and Voth (2012), and Obermeyer et al. (2014) for recent examples from policy analysis, economics, and health care. When there are multiple outcome variables of interest, there may exist unmeasured factors that influence a particular outcome while not impacting others. In order for these factors to affect the performed inference (and hence, to be confounders in the sense of VanderWeele and Shpitser (2013)), these factors must also impact the treatment assignment probabilities. Figure 3 demonstrates that these factors yield an aggregate impact on the assignment probabilities (U in the figure) despite affecting the outcomes differently. Controlling for the aggregate impact of unmeasured confounding on the assignment probabilities is sufficient for identifying the causal effect of the treatment on all of the outcome variables of interest, as these probabilities are themselves a minimally sufficient adjustment set (Rosenbaum and Rubin, 1983). The reader should keep in mind that u_{ij} truly reflects a dimension reduction of all unmeasured confounders to their relevant scalar component for impacting the assignment probabilities, and hence that this model for a sensitivity analysis does not limit the potential impact of unmeasured confounding on any of the outcome variables. Moving forward, we will refer to u_{ij} interchangeably as the “unmeasured confounder" and “unobserved covariate" for individual j in matched set i, as is conventional in sensitivity analyses following the model of Rosenbaum (2002a).

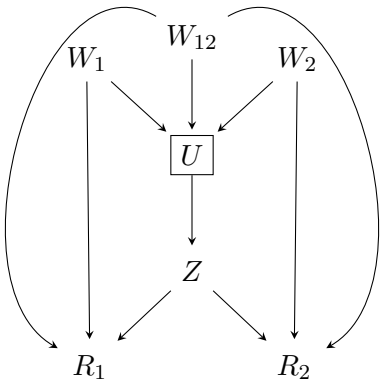

Figure 3: A Directed Acyclic Graph (DAG) showing how our method accounts for unmeasured confounding on multiple outcome variables by controlling for their joint effect on the treatment. W_1, W_2, W_{12} represent unmeasured factors which affect outcome R_1, outcome R_2, and both outcomes respectively. U is an aggregate measure of the impact of $\{W_1, W_2, W_{12}\}$ on the treatment assignment vector, Z. For any known value of U, only the direct causal pathway of the treatment, Z, on the outcome variables, R_1 and R_2, remains open if we condition on U (denoted by the square around U). Implicit in this diagram is that adjustment has been made for any observed pre-treatment confounders, X.

When conducting a sensitivity analysis with multiple outcomes, the unmeasured confounder which affects assignment probabilities in the worst-case manner for outcome k, $\mathbf{u}_k^*$, may not be worst-case for outcome k'; in fact, it may actually result in more *favorable* inference for outcome k'. As is noted in Rosenbaum and Silber (2009), naïvely combining the results of outcome-specific sensitivity analyses while accounting for multiple comparisons is unduly conservative precisely because of this: it allows the worst-case unmeasured confounder to affect the probabilities of assignment to treatment differently from one outcome to the next for the same test subject. Consequently, a uniform improvement in the power of a sensitivity analysis for testing the overall null hypothesis for any subset of outcomes could be attained by eliminating this logical inconsistency. As tests for the overall null hypothesis with respect to subsets of outcomes form the basis of multiple comparisons procedures such as closed testing (Marcus et al., 1976), hierarchical testing (Meinshausen, 2008), and the inheritance procedure (Goeman and Finos, 2012), such an advance would also uniformly improve the power of a sensitivity analysis for testing null hypotheses for *particular* outcomes while strongly controlling the familywise error rate.

The approaches for conducting a sensitivity analysis with a single outcome heavily utilize the fact that within each matched set, the search for the worst-case unmeasured confounder can be restricted to a readily enumerable set of binary vectors (Rosenbaum and Krieger, 1990). When testing whether the treatment has an effect on *at least one* of many outcome variables of interest, this is no longer the case. Thus, the potential gain in power cannot be actualized through simple extensions of existing methods. In this work, we present a new formulation of the required optimization problem as a quadratically constrained linear program which allows one to claim improved robustness to unmeasured confounding in an observational study with multiple outcomes when testing the overall null. This can, in turn, improve the reported robustness of individual level outcomes through its incorporation into certain sequential rejection procedures (Goeman and Solari, 2010). To illustrate these ideas, we now present an observational study on the impact of smoking on naphthalene levels in the body.

4.1.2. Motivating Example: Naphthalene Exposure in Smokers

Naphthalene is a simple polycyclic aromatic hydrocarbon (PAH) which has been linked to numerous adverse health outcomes. Exposure to excessive amounts of naphthalene can cause hemolysis (abnormal damage to or destruction of red blood cells in the body), which can in turn lead to hemolytic anemia (Todisco et al., 1991; Sanctucci and Shah, 2000). Further, naphthalene has been shown to be carcinogenic in animal studies (Hecht, 2002), prompting the International Agency for Research on Cancer (IARC) to label it as "possibly carcinogenic to humans" (IARC, 2002). Given the potential for adverse health outcomes from exposure to naphthalene, it is of interest to assess the impact of other sources of exposure to naphthalene on levels of naphthalene metabolites found in the body.

In the 2007-2008 National Health and Nutrition Examination Survey (NHANES), urinary concentrations of two monohydroxylated naphthalene metabolites, 1- and 2-naphthol (also known as α- and β-naphthol) were collected for 1706 patients from a representative sample of adults aged 20 and older in the United States. Through this study, we seek to address

the following question: after controlling for other sources of exposure and other relevant demographic variables, does smoking (one source of exposure to naphthalene) lead to elevated naphthalene metabolite levels in our study population? If this were the case, it would lend further credence to the belief that naphthalene is a useful biomarker for exposure to PAHs through inhalation (Nan et al., 2001; Hecht, 2002; Preuss et al., 2004), and it may serve to further highlight the health risks from smoking.

Through full matching (Hansen, 2004), 453 current smokers were placed into matched sets with 1253 non-current smokers who were similar on the basis of pre-treatment variables which, following the criterion for confounder selection of VanderWeele and Shpitser (2011), were deemed important to the decision to be a smoker or the outcomes; see Appendix C.1 for further details on the performed matching. Our two outcome variables were the urinary concentrations of 1- and 2-naphthol. Using an aligned rank test (Hodges and Lehmann, 1962) within the stratification yielded by our full match, we sought to determine whether there was evidence for smoking causing elevated levels for at least one of the two metabolites, and also whether smoking caused elevated metabolite levels for 1-naphthol and 2-naphthol considered individually. Assuming a multiplicative treatment effect model (additive on the log-scale), under no unmeasured confounding smoking was estimated to elevate urinary concentrations by a factor of 4.66 and 3.29 for 1- and 2-naphthol respectively using a Hodges-Lehman estimator (Hodges and Lehmann, 1963), with 95% confidence intervals of [4.00; 5.41] and [2.92; 3.69] attained by inverting a series of tests on the value of the multiplicative effect (Lehmann, 1963). Correcting for multiple comparisons using Holm-Bonferroni (Holm, 1979), the asymptotically separable algorithm of Gastwirth et al. (2000) applied individually to each metabolite yielded strong insensitivity to unmeasured confounding: the minimum and maximum of the two outcome-specific findings were below 0.025 and 0.05 respectively until a Γ of 7.78. This means that an unmeasured confounder would have to result in a difference in the odds of smoking for two individuals in the same matched set by a factor of 7.78 while nearly perfectly predicting naphthalene metabolite concentrations to render our results insignificant.

Based on these results, we can also attest to the robustness of a rejection of the *overall* null of no effect for either naphthalene metabolite: we have evidence for significance of at least one naphthalene metabolite at $\Gamma = 7.78$. As previously mentioned, this is conservative as using Holm-Bonferroni to combine individual sensitivity analyses allows for differing worst-case confounders for each outcome for the same individual. Naturally, the worst-case unmeasured confounder for 2-naphthol need not be the worst-case confounder for 1-naphthol. In fact, at $\Gamma = 7.78$ the worst-case $\mathbf{u}$ for 2-naphthol actually yields a significant result for 1-naphthol, and similarly the worst-case $\mathbf{u}$ for 1-naphthol makes our result for 2-naphthol significant. Through the methodology presented in this paper, it can be determined there is no vector of hidden covariates that simultaneously makes 1- and 2-naphthol insignificant at this level of unmeasured confounding. In fact, it takes a Γ of 10.22 to overturn the rejection of the overall null of no effect for either naphthalene metabolite. Thus $\Gamma = 7.78$ actually understates the robustness of a test of overall significance. Furthermore, we show in Section 4.5 that through a closed testing procedure we can actually claim robustness of the particular metabolites up until $\Gamma = 7.83$ for 1-naphthol and $\Gamma = 8.20$ for 2-naphthol, which are the same levels of robustness to unmeasured confounding that would have been arrived upon *without* controlling for multiple comparisons.

Section 4.2 provides notation for and a review of randomization inference and sensitivity analysis within a matched observational study. Section 4.3 introduces testing and sensitivity analysis for the overall null hypothesis when there are multiple outcomes. After highlighting the room for improvement relative to combining sensitivity analyses for each outcome, Section 4.4 formulates a quadratically constrained linear program which allows us to perform a sensitivity analysis for the overall null hypothesis while enforcing that for each outcome, the unmeasured confounder must be the same for each individual. Section 4.5 describes how our method can facilitate strong familywise error control for testing null hypotheses on particular outcomes through its incorporation into certain sequential rejection procedures. In Section 4.6, we present a simulation study demonstrating the potential gains in power of a sensitivity analysis on the overall null and on outcome-specific nulls using this procedure.

We return to our motivating example in Section 4.7, where we elucidate the improvements in reported robustness to unmeasured confounding attained through our procedure as they pertain to testing elevated naphthalene levels in smokers.

4.2. Notation for a Matched Observational Study

4.2.1. A Stratified Experiment with Multiple Outcomes

We now present notation for the idealized experiment targeted by matching algorithms wherein each treated unit is placed in a matched set with one or more control units. This framework can be trivially extended to dealing with strata resulting from full matching, such as the one presented in Section 4.1.2; see Rosenbaum (2002a, Section 4, Problem 12) for details. Suppose there are I independent strata, the i^{th} of which contains $n_i \geq 2$ individuals, that were formed on the basis of pre-treatment covariates. In each stratum, 1 individual receives the treatment and $n_i - 1$ individuals receive the control. There are K outcome variables collected for each individual. For each outcome k, individual j in stratum i has two potential outcomes: one under treatment, r_{Tijk}, and one under control, r_{Cijk}; let $\mathbf{r}_{Tij}$ and $\mathbf{r}_{Cij}$ be the K-dimensional vector of potential outcomes for this individual. The observed response vector for each individual is $\mathbf{R}_{ij} = \mathbf{r}_{Tij}Z_{ij} + \mathbf{r}_{Cij}(1 - Z_{ij})$, where Z_{ij} is an indicator variable that takes the value 1 if individual j in stratum i is assigned to the treatment; see, for example, Neyman (1923) and Rubin (1974). Each individual has a vector of observed covariates $\mathbf{x}_{ij}$ and an unobserved covariate u_{ij}.

There are $N = \sum_{i=1}^{I} n_i$ total individuals in the study. Let $\mathbf{Z} = [Z_{11}, Z_{12}, ...,, Z_{In_I}]^T$ be the binary vector of treatment assignments, and let $\mathbf{R}, \mathbf{r}_T$, and $\mathbf{r}_C$ be the $N \times K$ matrices of observed responses and potential outcomes under treatment and control. Let Ω be the set of $\prod_{i=1}^{I} n_i$ possible values of $\mathbf{Z}$ under the given stratification. In randomization inference for a randomized experiment, randomness is modeled solely through the assignment to treatment or to control (Fisher, 1935). Quantities dependent on $\mathbf{Z}$, such as the observed outcomes $\mathbf{R}$, are random, while $\mathbf{r}_{Tij}, \mathbf{r}_{Cij}, \mathbf{x}_{ij}$, and u_{ij} are fixed across randomizations. Let

$\mathcal{F}$ be the set of such fixed quantities. For a randomized experiment adhering to this design $\mathbb{P}(Z_{ij} = 1|\mathcal{F}, \mathbf{Z} \in \Omega) = 1/n_i$ and $\mathbb{P}(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathbf{Z} \in \Omega) = 1/|\Omega|$, where $|A|$ denotes the number of elements in a finite set A.

4.2.2. Randomization Inference and Sensitivity Analysis

For each outcome k, we consider hypotheses of the form $H_k : f_{Tk}(r_{Tijk}) = f_{Ck}(r_{Cijk})\ \forall i, j$ for specified functions $f_{Tk}(\cdot)$ and $f_{Ck}(\cdot)$. For example, Fisher's sharp null of no effect can be tested through $f_{Tk}(r_{Tijk}) = r_{Tijk}$ and $f_{Ck}(r_{Cijk}) = r_{Cijk}$, and a test of an additive treatment effect τ_k can be tested by setting $f_{Tk}(r_{Tijk}) = r_{Tijk}$ and $f_{Ck}(r_{Cijk}) = r_{Cijk} + \tau_k$. While tests for Neyman's weak null of no average treatment effect cannot be accommodated within the framework that follows, other choices of $f_{Tk}(\cdot)$ and $f_{Ck}(\cdot)$ can yield tests allowing for subject-specific causal effects such as tests of effect modification, dilated treatment effects, displacement effects, tobit effects, and attributable effects; see Rosenbaum (2002a, Section 5) and Rosenbaum (2010, Sections 2.4-2.5) for an overview.

From our data alone we observe $F_{ijk} = f_{Tk}(r_{Tijk})Z_{ij} + f_{Ck}(r_{Cijk})(1 - Z_{ij})$; let $\mathbf{F}_k = [F_{11k}, ..., F_{In_Ik}]$. Under H_k, the vectors $\mathbf{f}_{Ck} = [f_{Ck}(r_{C11k}), ..., f_{Ck}(r_{CIn_Ik})]$ and $\mathbf{f}_{Tk} = [f_{Tk}(r_{T11k}), ..., f_{Tk}(r_{TIn_Ik})]$ are known to be equal, and hence are entirely specified. Further, they are constant across randomizations as they are known functions of the potential outcomes. Hence, under the null $\mathbf{F}_k = \mathbf{f}_{Tk} = \mathbf{f}_{Ck} \in \mathcal{F}$, which in turn allows us to use randomization inference to test H_k. Specifically, under H_k and under the stratified experiment discussed in Section 4.2.1 the null distribution of a test statistic $t_k(\mathbf{Z}, \mathbf{F}_k)$ can be written as:

$$\mathbb{P}\{t_k(\mathbf{Z}, \mathbf{F}_k) \geq a|\mathcal{F}, \mathbf{Z} \in \Omega; H_k\} = \frac{|\mathbf{z} \in \Omega : t_k(\mathbf{z}, \mathbf{f}_{Ck}) \geq a|}{|\Omega|}, \tag{4.1}$$

where we use $\mathbf{f}_{Ck}$ in the right-hand side to emphasize that this distribution is known under the null.

The distribution of $t_k(\mathbf{Z}, \mathbf{F}_k)$ in (4.1) is appropriate if the observed data truly resulted from the randomized experiment described in Section 4.2.1. In an observational study

employing matching, we aim to replicate this idealized randomized experiment by creating strata wherein individuals are similar on the basis of their observed covariates, $\mathbf{x}_{ij}$ (Ming and Rosenbaum, 2000; Hansen, 2004; Stuart, 2010). While this seeks to control for observed confounders, individuals placed in a given stratum i may be different on the basis of the unobserved covariate u_{ij}. If this u_{ij} is influential for the assignment of treatments and the response, the distribution in (4.1) may yield highly misleading inferences.

We follow the model for a sensitivity analysis discussed in Rosenbaum (2002a, Section 4), which states that failure to account for unobserved covariates may result in biased treatment assignments within a stratum. This model can be parameterized by a number $\Gamma = \exp(\gamma) \geq 1$ which bounds the extent to which the odds ratio of assignment can vary between two individuals who are in the same matched stratum. Under this formulation, the probability of a given allocation of treatment and control within the stratification under consideration can be stated in the form $\mathbb{P}(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathbf{Z} \in \Omega) = \exp(\gamma \mathbf{z}^T \mathbf{u}) / \sum_{\mathbf{b} \in \Omega} \exp(\gamma \mathbf{b}^T \mathbf{u})$, where $\mathbf{u} = [u_{11}, u_{12}, ..., u_{I,n_i}] \in [0,1]^N =: \mathcal{U}$ is a vector of unmeasured confounders for the individuals in the study. Note that $\Gamma = 1$ corresponds to the randomization distribution discussed in Section 4.2.1, while for $\Gamma > 1$ the resulting distribution differs from that of a randomized experiment, with Γ controlling the extent of this departure.

We consider test statistics of the form $t_k(\mathbf{Z}, \mathbf{F}_k) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} Z_{ij} q_{ijk}$, where q_{ijk} are functions of $\mathbf{F}_k$. Under H_k these values become functions of $\mathbf{f}_{Ck}$, and hence are known quantities fixed across randomizations. Let $\mathbf{q}_k = [q_{11k}, ..., q_{In_Ik}]$, and let $\mathbf{q}_{ik} = [q_{i1k}, ..., q_{in_ik}]$. Many commonly employed statistics can be written in this form. For example, suppose we are testing Fisher's sharp null, so that $F_{ijk} = R_{ijk}$, within the block-randomized experiment described in Section 4.2.1. Setting $q_{ijk} = \sum_{j' \neq j} (F_{ijk} - F_{ij'k})/(I(n_i - 1))$, $t_k(\mathbf{Z}, \mathbf{F}_k)$ is the mean over the I matched sets of the average treated-minus-control difference in each matched set for outcome k. In the case of a matched pairs design, $n_i = 2 \ \ \forall i$, this yields the paired permutation t-test. If q_{ijk} are the ranks of the aligned response $F_{ijk} - \sum_{j'=1}^{n_i} F_{ij'k}/n_i$ from 1 to N, then a test on $t_k(\mathbf{Z}, \mathbf{F}_k)$ yields the aligned rank test of Hodges and Lehmann (1962).

To recover Wilcoxon's signed rank statistic for a matched pairs design, let d_{ik} be the ranks of $|F_{i1k} - F_{i2k}|$ from 1 to I, and let $q_{ijk} = d_{ik}1\{F_{ijk} > F_{ij'k}\}$. See Rosenbaum (2002a) for additional examples and further discussion.

For any given value of $\Gamma \geq 1$, a sensitivity analysis proceeds by finding the allocation of the unmeasured confounder $\mathbf{u}^*$ which maximizes the p-value for the hypothesis test being conducted. While not explicitly noted, this worst-case unmeasured confounder can vary with the value of Γ under investigation. One then finds the smallest value of Γ such that the conclusions of the study would be altered (i.e., such that the conclusion of the hypothesis test would change from rejecting to failing to reject the null hypothesis). The more robust a given study is to unmeasured confounding, the larger the value of Γ must be to alter its findings. Under mild regularity conditions on $\mathbf{q}_k$, the distribution under the null of $t_k(\mathbf{Z}, \mathbf{F}_k)$ converges to that of a normal random variable as $I \to \infty$ for the worst-case confounder $\mathbf{u}^*$ at any Γ. An example of regularity conditions on the constants q_{ijk} is that the Lindeberg condition holds for the random variables $B_{ik} := \sum_{j=1}^{n_i} Z_{ij} q_{ijk}$ (Lehmann, 2004, Theorem A.1.1). While the value of Γ itself does not affect the limiting distribution, it does influence the rate at which this limit is reached as larger values of Γ allow for larger discrepancies in the assignment probabilities within a matched set. Under asymptotic normality, large sample bounds on the tail probability can instead be expressed in terms of corresponding bounds on standardized deviates.

For further discussion of sensitivity analyses, including illustrations and alternate models, see Cornfield et al. (1959), Marcus (1997), Imbens (2003), Yu and Gastwirth (2005), Wang and Krieger (2006), Egleston et al. (2009), Hosman et al. (2010), VanderWeele and Arah (2011), Zubizarreta et al. (2013), Liu et al. (2013) and Ding and Vanderweele (2014).

4.3. Sensitivity Analysis for Overall Significance

4.3.1. Testing the Overall Null Hypothesis

We begin with notation for the truth of the null hypotheses on all K outcomes; extensions of notation to dealing with subsets of outcomes, which will in turn facilitate strong familywise error control for testing individual outcomes, will be made in Section 4.5. There are K hypotheses, $H_1, ..., H_K$, and we are interested in testing the overall truth of the hypotheses $\{H_1, .., H_K\}$ while strongly controlling the familywise error rate at level α for a range of Γ.

$$\mathbf{H_o} : \bigwedge_{k=1}^{K} H_k$$
$$\mathbf{H_a} : \bigvee_{k=1}^{K} H_k^c$$

We will refer to a test of $\mathbf{H_o}$ as a test of the *overall null*. Moving forward, we assume each individual hypothesis H_k has an associated test statistic $t_k(\mathbf{Z}, \mathbf{F}_k)$ of the form discussed in Section 4.2.2.

4.3.2. Combining Individual Sensitivity Analyses is Conservative

A simple approach for conducting a sensitivity analysis at a given Γ would be to separately find the worst-case p-value for each hypothesis test, call it P_k^* with corresponding allocation of worst-case confounder $\mathbf{u}_k^*$, and suggest through the use of a Bonferroni correction that at least one hypothesis is false if $\min_k P_k^* \leq \alpha/K$. This trivially controls familywise error rate at α as desired; however, as is noted in Rosenbaum and Silber (2009, Section 4.5), this approach is conservative as the worst-case p-value for hypothesis test k may be found at a different allocation of the unmeasured confounder as that of hypothesis test $k' \neq k$ for $k, k' \in \{1, ..., K\}$ (i.e., $\mathbf{u}_k^* \neq \mathbf{u}_{k'}^*$). In other words, the biased treatment assignment probabilities caused by unmeasured confounding that yield the worst-case inference for outcome k and outcome k' need not be the same. This can be better understood in light of the following

well known minimax inequality (for instance, Karlin, 1992, Lemma 1.3.1)

$$\min_{k\in\{1,..,K\}} \max_{\mathbf{u}\in\mathcal{U}} P_{k,\mathbf{u}} \geq \max_{\mathbf{u}\in\mathcal{U}} \min_{k\in\{1,..,K\}} P_{k,\mathbf{u}}. \tag{4.2}$$

Combining the results of K separate hypothesis tests and Bonferroni correcting corresponds to the left-hand side of (4.2). Strict inequality is possible in (4.2): it could be the case that $\min_k \max_{\mathbf{u}\in\mathcal{U}} P_k > \alpha/K$, meaning that we would fail to reject the overall null hypothesis if we conducted sensitivity analyses separately for each k and then Bonferroni corrected, while in reality $\max_{\mathbf{u}\in\mathcal{U}} \min_k P_k \leq \alpha/K$, such that we should have rejected the overall null. This would occur if for each k there exists a $\mathbf{u}_k^* \in \mathcal{U}$ such that H_k is not rejected, yet there does not exist a *single* $\mathbf{u}^* \in \mathcal{U}$ for which all H_k are simultaneously not rejected.

A uniform improvement over combining individual sensitivity analyses could be achieved by a procedure which directly solved for the right-hand side of (4.2). Such a procedure cannot be derived by extending existing methods for conducting individual level sensitivity analyses, as these methods rely upon the fact that the search for a worst-case confounder can be restricted to vectors in $\mathcal{U}^+$ or $\mathcal{U}^-$ for any particular hypothesis k. Unfortunately, it is not the case that vector $\mathbf{u}^*$ which achieves
$\max_{\mathbf{u}\in\mathcal{U}} \min_{k\in\{1,..,K\}} P_{k,\mathbf{u}}$ lies within an easily enumerated set of vertices of $\mathcal{U}$; in fact, the solution need not even lie at a vertex. To exploit this potential improvement, a new formulation of the required optimization problem that allows for solutions in all of $\mathcal{U}$ is thus required.

4.4. Improving Power through Quadratically Constrained Linear Programming

In this section, we assume the individual level hypotheses H_k have two-sided alternatives; simple extensions to the one-sided case are discussed in Appendix C.2. Using a normal approximation, we can equivalently express our problem as minimizing over $\mathcal{U}$ the maximal

squared deviate over the K hypotheses in question:

$$\min_{\mathbf{u}\in\mathcal{U}} \max_{k\in\{1,..,K\}} \frac{(t_k - \mu_{k,\mathbf{u}})^2}{\sigma^2_{k,\mathbf{u}}}, \tag{4.3}$$

where t_k is the observed value of the statistic $t_k(\mathbf{Z}, \mathbf{F}_k)$, and $\mu_{k,\mathbf{u}} = \mathbb{E}_{\Gamma,\mathbf{u}}[\mathbf{Z}^T\mathbf{q}_k|\mathcal{F}, \mathbf{Z} \in \Omega]$ and $\sigma^2_{k,\mathbf{u}} = \mathbf{Var}_{\Gamma,\mathbf{u}}(\mathbf{Z}^T\mathbf{q}_k|\mathcal{F}, \mathbf{Z} \in \Omega)$ are the means and variances of the test statistic $t_k(\mathbf{Z}, \mathbf{F}_k)$ with a given value of Γ and vector $\mathbf{u}$ under the permutation distribution given by (4.1). Under a normal approximation for $t_k(\mathbf{Z}, \mathbf{F}_k)$, the squared deviate follows a χ^2_1 distribution. Hence, a determination of whether or not we can reject at least one null hypothesis can be made by checking whether or not the solution to (4.3) is greater than or equal to $\chi^2_{1,1-\alpha/K}$, where $\chi^2_{1,1-\alpha/K}$ is the $1 - \alpha/K$ quantile of a χ^2_1 distribution.

Moving forward, all expectations and variances are taken with respect to the distribution in (4.1), i.e. under the truth of the null hypothesis H_k for each k, and are conditional on $\mathcal{F}$ and $\mathbf{Z} \in \Omega$; this is omitted for notational ease. Let $\varrho_{ij} = \exp(\gamma u_{ij})/\sum_{j'=1}^{n_i} \exp(\gamma u_{ij'}) = \mathbb{P}(Z_{ij} = 1|\mathcal{F}, \mathbf{Z} \in \Omega)$. Let $\boldsymbol{\varrho}_i = [\varrho_{i1}, .., \varrho_{in_i}]$, and let $\boldsymbol{\varrho} = [\varrho_{11}, .., \varrho_{In_I}]$. Note that we can express our test statistics as the sums of stratum-wise contributions, $t_k(\mathbf{Z}, \mathbf{F}_k) = \sum_{i=1}^{I} B_{ik}$ where $B_{ik} := \sum_{j=1}^{n_i} Z_{ij}q_{ijk}$. The expectation and variance of the contribution from stratum i, B_{ik}, can be written as

$$\begin{aligned}\mathbb{E}[B_{ik}; \boldsymbol{\varrho}] &= \boldsymbol{\varrho}_i^T \mathbf{q}_{ik} \\ \mathbf{Var}(B_{ik}; \boldsymbol{\varrho}) &= \boldsymbol{\varrho}_i^T \mathbf{q}_{ik}^2 - (\boldsymbol{\varrho}_i^T \mathbf{q}_{ik})^2,\end{aligned}$$

where the simplified form of $\mathbf{Var}(B_{ik}; \boldsymbol{\varrho})$ comes from the constraint that $\sum_{j=1}^{n_i} Z_{ij} = 1 \; \forall i$.

For a given $\boldsymbol{\varrho}$, we can reject the null hypothesis for a two sided alternative at level α/K if $(t_k - \mathbb{E}[t_k(\mathbf{Z}, \mathbf{F}_k); \boldsymbol{\varrho}])^2/\mathbf{Var}(t_k(\mathbf{Z}, \mathbf{F}_k); \boldsymbol{\varrho}) \geq \chi^2_{1,1-\alpha/K}$, where $\mathbb{E}[t_k(\mathbf{Z}, \mathbf{F}_k); \boldsymbol{\varrho}] = \sum_{i=1}^{I} \mathbb{E}[B_{ik}; \boldsymbol{\varrho}]$, and $\mathbf{Var}(t_k(\mathbf{Z}, \mathbf{F}_k); \boldsymbol{\varrho}) = \sum_{i=1}^{I} \mathbf{Var}(B_{ik}; \boldsymbol{\varrho})$ due to independence between strata. This is equivalent to rejecting if $\zeta_k(\boldsymbol{\varrho}) := (t_k - \mathbb{E}[t_k(\mathbf{Z}, \mathbf{F}_k); \boldsymbol{\varrho}])^2 - \chi^2_{1,1-\alpha/K}\mathbf{Var}(t_k(\mathbf{Z}, \mathbf{F}_k); \boldsymbol{\varrho}) \geq 0$. If we can determine that $\zeta_k(\boldsymbol{\varrho}) \geq 0$ for all feasible values of $\boldsymbol{\varrho}$ at a given value of Γ, we can then

say that we have rejected the null at level of unmeasured confounding Γ; otherwise, we fail to reject.

Consider the following optimization problem:

$$\begin{aligned} &\underset{\varrho_{ij}, s_i}{\text{minimize}} \;\; \zeta_k(\boldsymbol{\varrho}) && (H_k) \\ &\text{subject to} \;\; \sum_{j=1}^{n_i} \varrho_{ij} = 1 \;\; \forall i \\ &\qquad\qquad s_i \leq \varrho_{ij} \leq \Gamma s_i \;\; \forall i, j \\ &\qquad\qquad \varrho_{ij} \geq 0 \;\; \forall i, j \end{aligned}$$

The variables s_i stem from an application of a Charnes-Cooper transformation, $s_i = 1/\sum_{j'=1}^{n_i} \exp(\gamma u_{ij'})$ (Charnes and Cooper, 1962), and allow us to incorporate the restrictions on the allowable departure from pure randomization, $1 \leq \exp(\gamma u_{ij}) \leq \Gamma \;\; \forall i, j$, in terms of the probabilities themselves.

Problem (P2) is a quadratic program, which can be readily solved using a host of free and commercially available solvers; however, solving this problem merely results in a sensitivity analysis for a *particular* hypothesis H_k, rather than one of the overall null $\wedge H_k$. Towards this end, define $\zeta(\boldsymbol{\varrho}) = \max\{\zeta_1(\boldsymbol{\varrho}), ..., \zeta_K(\boldsymbol{\varrho})\}$. We can now pose our problem as finding $\min_{\boldsymbol{\varrho}} \zeta(\boldsymbol{\varrho})$ subject to constraints on $\boldsymbol{\varrho}$ imposed by Γ. This optimization can be performed through incorporating an auxiliary variable y and solving the following quadratically constrained

linear program:

$$\begin{aligned} \underset{y,\varrho_{ij},s_i}{\text{minimize}} \;\; & y && (\wedge H_k) \\ \text{subject to} \;\; & y \geq \zeta_k(\boldsymbol{\varrho}) \;\; \forall k \\ & \sum_{j=1}^{n_i} \varrho_{ij} = 1 \;\; \forall i \\ & s_i \leq \varrho_{ij} \leq \Gamma s_i \;\; \forall i, j \\ & \varrho_{ij} \geq 0 \;\; \forall i, j \end{aligned}$$

The auxiliary variable y is forced to be larger than $\zeta_k(\boldsymbol{\varrho})$ for all k, and by minimizing over y the optimization problem searches for the feasible value of $\boldsymbol{\varrho}$ that allows for y to become as small as possible, hence minimizing the maximum value as desired. This is a commonly employed device for solving minimax problems; see, for example, Charalambous and Conn (1978). To determine whether or not we can reject at least one null hypothesis, we simply check whether the optimal value $y^* \geq 0$. If it is, we can reject at least one null hypothesis; otherwise, we cannot. Quadratically constrained linear programs can be solved using many available solvers; we provide an implementation using the `R` interface to `Gurobi`, a commercial solver which is freely available for academic use. Henceforth, we will refer to this procedure for conducting a sensitivity analysis the overall null with K outcomes as the "minimax" procedure (for minimizing the maximum squared deviate).

4.5. Familywise Error Control for Individual Null Hypotheses

By addressing the right-hand side of (4.2), the minimax procedure provides a sensitivity analysis for the overall null hypothesis that uniformly dominates combining individual sensitivity analyses. In this section, we discuss how the minimax procedure can be used with sequential rejection procedures (Goeman and Solari, 2010) which progress through testing the overall null for a sequence of subsets of outcomes (henceforth referred to as intersection nulls) to provide uniform improvements in power for testing hypotheses on *particular*

outcome variables. Sequential rejection procedures of this sort include closed testing (Marcus et al., 1976), hierarchical testing (Meinshausen, 2008), and the inheritance procedure (Goeman and Finos, 2012). These procedures have appealing properties for conducting a sensitivity analysis, often allowing researchers to claim improved robustness of a study's findings against unmeasured confounding; see Rosenbaum and Silber (2009) for a discussion of this fact as it relates to closed testing procedures.

We now introduce notation for the class of sequential rejection procedures which can be used in conjunction with our method, i.e. those for which each step involves testing the truth of an intersection null hypothesis for a subset of the K outcome variables. There are L intersection null hypotheses ordered from 1,...,L, the ℓ^{th} of which, $\mathbf{H_o}_\ell$, pertains to the null hypothesis being true for all outcomes in the subset $\mathcal{K}_\ell \subseteq \{1, ..., K\}$. That is, $\mathbf{H_o}_\ell = \bigwedge_{k \in \mathcal{K}_\ell} H_{k\ell}$. $|\mathcal{K}_\ell| \leq K$ is the number of outcomes being tested in the ℓ^{th} subset; $|\mathcal{K}_\ell| = 1$ then corresponds to a test of a particular outcome. Let $\mathcal{H}$ be the set of these L intersection null hypotheses, $\mathcal{H} = \{\mathbf{H_o}1, ..., \mathbf{H_o}L\}$.

Following Goeman and Solari (2010), let $\mathcal{R}_a \subseteq \mathcal{H}$ be the collection of intersection nulls rejected after step a of the sequential rejection procedure, and let $\mathcal{N}(\mathcal{R}_a)$ be the set of intersection nulls that can now be rejected in step $a + 1$ if all elements of $\mathcal{R}_a$ have been rejected by step a. The sequential rejection procedure can then be defined by

$$\mathcal{R}_0 = \emptyset$$
$$\mathcal{R}_{a+1} = \mathcal{R}_a \cup \mathcal{N}(\mathcal{R}_a),$$

and is repeated until convergence (i.e., until $\mathcal{R}_{a+1} = \mathcal{R}_a$). Goeman and Solari (2010) show that sequential rejection procedures strongly control the familywise error rate at α under the conditions (1) the procedure controls the familywise error at α for the so-called *critical case* in which procedure has rejected all of the false overall null hypotheses and none of the true overall nulls and (2) no false rejections in the critical case implies no false rejections in situations with fewer rejections than the critical case.

Closed testing, hierarchical testing, and the inheritance procedure can all be recovered through specific choices of $\mathcal{N}(\cdot)$ that provably adhere to these conditions. Testing the intersection nulls $\mathbf{H_{o\ell}}$ for any ℓ at level of unmeasured confounding Γ as required by these procedures can be performed using the minimax procedure of Section 4.4, which through inequality (4.2) provides improved power for each subset tested.

To illustrate, suppose one is interested in using a closed testing procedure to conduct a sensitivity analysis with $K = 2$ outcomes; this is the procedure used for multiple testing in our motivating example. In this case, $L = 3$, $\mathcal{K}_1 = \{1, 2\}$, $\mathcal{K}_2 = \{1\}$, $\mathcal{K}_3 = \{2\}$. The function $\mathcal{N}(\cdot)$ then takes on the following form:

$$\mathcal{N}(\emptyset) = \begin{cases} \mathbf{H_{o1}} & \text{if reject } H_1 \wedge H_2 \text{ at level } \alpha \\ \emptyset & \text{otherwise} \end{cases}$$

$$\mathcal{N}(\mathbf{H_{o1}}) = \begin{cases} \{\mathbf{H_{o1}}, \mathbf{H_{o2}}, \mathbf{H_{o3}}\} & \text{if } H_1 \text{ and } H_2 \text{ each reject individually at level } \alpha \\ \{\mathbf{H_{o1}}, \mathbf{H_{o2}}\} & \text{if only } H_1 \text{ rejects at level } \alpha \\ \{\mathbf{H_{o1}}, \mathbf{H_{o3}}\} & \text{if only } H_2 \text{ rejects at level } \alpha \\ \{\mathbf{H_{o1}}\} & \text{otherwise,} \end{cases}$$

and $\mathcal{N}(A) = A$ if $A \neq \emptyset$ and $A \neq \mathbf{H_{o1}}$. In this example, the test of $\mathbf{H_{o1}}$ can be performed using the minimax procedure with a test that is locally level α; the tests of $\mathbf{H_{o2}}$ and $\mathbf{H_{o3}}$ only involve one outcome and thus can be conducted through the usual methods for a sensitivity analysis which, by the closure principle, can be performed locally at level α while strongly controlling the familywise error rate.

4.6. Simulation Study: Gains in Power of a Sensitivity Analysis

4.6.1. Overall Null Hypothesis

Through the minimax procedure, we arrive at a uniform improvement for testing the overall null relative to combining the results of individual sensitivity analyses. In this section, we

present a simulation study to demonstrate the potential gains in power for testing the overall null. In each of 24 simulation settings, we simulate 10,000 data sets with $I = 250$ pairs and $K = 5$ outcome variables of interest. The vector of treated-minus-control paired differences $\mathbf{D}_i$ are simulated iid from a multivariate normal with mean vector $\boldsymbol{\tau}$ and covariance matrix Σ. For each outcome, we use an M-statistic of the type favored by Huber (1981), $t_k(\mathbf{Z}, \mathbf{F}_k) = \sum_{i=1}^{I} \psi(D_{ik}/s_k)$, to conduct inference, where s_k is the median of $|D_{ik}|$ across individuals i and $\psi(y) = \text{sign}(y)\min(|y|, 2.5)$. See Maritz (1979) for a discussion of randomization inference for M-statistics, and see Rosenbaum (2007, 2013, 2014) for various aspects of sensitivity analyses for M-statistics.

In evaluating these two procedures, we assume as is advocated in Rosenbaum (2004, 2007) that unbeknownst to the practitioner the paired data at hand truly arose from a stratified randomized experiment (i.e., $\Gamma = 1$). Hence, using a standard randomization test without assuming unmeasured confounding would provide honest type I error control. The practitioner, blind to this, would like to not only perform inference under the assumption of no unmeasured confounding, but also assess the robustness of the study's findings to unobserved biases of varying severity.

Our 24 simulation settings are the 8 possible combinations of the following mean and covariance vectors, each tested at $\Gamma = 1.25, 1.5$ and 1.75:

1. $\boldsymbol{\tau}^{(1)} = [0.25, 0.25, 0.25, 0.25, 0.25]$; $\boldsymbol{\tau}^{(2)} = [0.25, 0.25, 0.25, 0.25, 0]$;
   $\boldsymbol{\tau}^{(3)} = [0.3, 0.3, 0, 0, 0]$; $\boldsymbol{\tau}^{(4)} = [0.3, 0, 0, 0, 0]$

2. $\Sigma^{(1)} = \text{Diag}(1)$; $\Sigma^{(2)}_{ij} = 1$ if $i = j$, $\Sigma^{(2)}_{ij} = 0.5$ otherwise.

All hypothesis tests are of Fisher's sharp null, and are conducted with two-sided alternatives at $\alpha = 0.05$. Table 6 displays the probabilities of (correctly) rejecting the overall null of no effect for any of the outcomes. The first column contains the probabilities of rejection when combining the results of individual sensitivity analyses, while the second contains these probabilities for the minimax procedure. The relative improvement through the minimax

Table 6: Power of a sensitivity analysis for the overall null.

Gamma	Moments	Separate	Minimax
$\Gamma = 1.25$	$\boldsymbol{\tau}^{(1)}, \Sigma^{(1)}$	0.94	0.99
	$\boldsymbol{\tau}^{(1)}, \Sigma^{(2)}$	0.77	0.80
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(1)}$	0.89	0.96
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(2)}$	0.73	0.77
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(1)}$	0.92	0.96
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(2)}$	0.85	0.87
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(1)}$	0.72	0.72
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(2)}$	0.71	0.72
$\Gamma = 1.5$	$\boldsymbol{\tau}^{(1)}, \Sigma^{(1)}$	0.34	0.78
	$\boldsymbol{\tau}^{(1)}, \Sigma^{(2)}$	0.25	0.33
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(1)}$	0.28	0.66
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(2)}$	0.21	0.28
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(1)}$	0.45	0.65
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(2)}$	0.39	0.45
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(1)}$	0.26	0.26
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(2)}$	0.25	0.25
$\Gamma = 1.75$	$\boldsymbol{\tau}^{(1)}, \Sigma^{(1)}$	0.04	0.36
	$\boldsymbol{\tau}^{(1)}, \Sigma^{(2)}$	0.03	0.06
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(1)}$	0.03	0.23
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(2)}$	0.03	0.05
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(1)}$	0.09	0.24
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(2)}$	0.09	0.12
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(1)}$	0.05	0.05
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(2)}$	0.04	0.04

procedure can be quite substantial when the paired differences are independent across outcomes ($\Sigma^{(1)}$), while more modest improvements are attained when the paired differences are positively correlated ($\Sigma^{(2)}$). With positively correlated differences across outcomes, the worst-case unmeasured confounder for a particular outcome begins to align more closely with the worst-case unmeasured confounder for the other outcomes, while for independent paired differences this often is not the case. For both independent and correlated paired differences, gains are also more substantial when there are 5 or 4 nonzero treatment effects ($\boldsymbol{\tau}^{(1)}$ and $\boldsymbol{\tau}^{(2)}$) versus 2 larger nonzero effects ($\boldsymbol{\tau}^{(3)}$), and with only one large nonzero effects ($\tau^{(4)}$) the two methods tend to coincide. With fewer nonzero effects, the significance of the overall null at a given level of unmeasured confounding depends on the pattern of paired

differences in a small number of outcomes, such that even if the worst-case unmeasured confounder for an outcome with a nonzero effect actually improves the squared deviate for an outcome with zero effect it is unlikely to elevate said deviate to a level of significance.

Naturally, the probabilities of rejection decrease as Γ increases for each combination of mean vector and covariance matrix. We also note that as Γ increases, the gains from using the minimax procedure also increase . For example, with combination $\boldsymbol{\tau}^{(2)}, \Sigma^{(1)}$ the powers of the combined approach versus the minimax approach are 0.89 and 0.96 at $\Gamma = 1.25$, and are 0.28 versus 0.66 at $\Gamma = 1.5$. These simulations indicate that conducting a sensitivity analysis for the overall null by minimizing the maximum squared deviate allows for substantial and clinically relevant gains in the power of a sensitivity analysis. Additionally, the computational burden of the required optimization problem was minimal in these simulations: across all 24 simulation settings, the average computation time on a desktop computer with a 3.40 GHz processor and 16.0 GB RAM was 0.12 seconds.

4.6.2. Individual Hypotheses

As discussed in Section 4.5, the benefits of our procedure extend beyond testing the overall null, and can in fact yield improved power for a sensitivity analysis on hypotheses for individual outcomes. To illustrate this fact, we present a simulation study assessing the individual-level power of a sensitivity analysis for each of $K = 3$ outcomes. We use a closed testing procedure in order to test hypotheses on individual outcomes. Briefly, the closed testing principle states that if there are K hypotheses $H_1, ..., H_K$ that are of interest, we can reject any particular hypothesis H_k with familywise error control at α if all intersections of hypotheses including H_k can be rejected with tests that are individually level α. For example, with three outcomes we can reject H_1 if we can reject $H_1 \wedge H_2 \wedge H_3$, $H_1 \wedge H_2$, $H_1 \wedge H_3$, and H_1 with tests that are locally level α. When combining the results of individual sensitivity analyses, this equates to the Holm-Bonferroni procedure. When using the minimax procedure for closed testing, one instead solves problem $(\wedge H_k)$ for each intersection hypothesis.

Table 7: Power of closed testing for individual nulls.

		Separate				Minimax			
Gamma	Moments	H_1	H_2	H_3	$\wedge H_k$	H_1	H_2	H_3	$\wedge H_k$
$\Gamma = 1.25$	$\boldsymbol{\tau}^{(1)}, \Sigma^{(1)}$	0.27	0.40	0.54	0.74	0.33	0.46	0.60	0.84
	$\boldsymbol{\tau}^{(1)}, \Sigma^{(2)}$	0.29	0.40	0.53	0.62	0.31	0.43	0.56	0.65
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(1)}$	0.65	0.86	0.96	0.99	0.68	0.88	0.97	1.00
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(2)}$	0.65	0.85	0.95	0.97	0.66	0.86	0.96	0.97
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(1)}$	0.32	0.59	0.95	0.97	0.35	0.63	0.97	0.99
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(2)}$	0.34	0.58	0.94	0.95	0.35	0.60	0.95	0.95
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(1)}$	0.09	0.27	0.94	0.95	0.11	0.29	0.95	0.97
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(2)}$	0.11	0.27	0.93	0.94	0.11	0.28	0.94	0.94
$\Gamma = 1.375$	$\boldsymbol{\tau}^{(1)}, \Sigma^{(1)}$	0.09	0.16	0.27	0.41	0.14	0.22	0.34	0.61
	$\boldsymbol{\tau}^{(1)}, \Sigma^{(2)}$	0.11	0.18	0.27	0.35	0.13	0.20	0.30	0.39
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(1)}$	0.37	0.63	0.85	0.94	0.42	0.70	0.90	0.99
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(2)}$	0.39	0.62	0.84	0.87	0.41	0.65	0.85	0.89
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(1)}$	0.12	0.31	0.83	0.87	0.16	0.37	0.88	0.95
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(2)}$	0.14	0.32	0.82	0.83	0.16	0.35	0.83	0.84
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(1)}$	0.02	0.10	0.81	0.83	0.03	0.12	0.85	0.89
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(2)}$	0.03	0.11	0.82	0.82	0.03	0.12	0.82	0.82
$\Gamma = 1.5$	$\boldsymbol{\tau}^{(1)}, \Sigma^{(1)}$	0.03	0.06	0.11	0.18	0.05	0.09	0.16	0.36
	$\boldsymbol{\tau}^{(1)}, \Sigma^{(2)}$	0.03	0.06	0.12	0.16	0.05	0.08	0.14	0.19
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(1)}$	0.16	0.38	0.64	0.77	0.22	0.48	0.76	0.95
	$\boldsymbol{\tau}^{(2)}, \Sigma^{(2)}$	0.18	0.38	0.64	0.69	0.20	0.42	0.68	0.74
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(1)}$	0.04	0.13	0.62	0.66	0.06	0.18	0.71	0.84
	$\boldsymbol{\tau}^{(3)}, \Sigma^{(2)}$	0.04	0.14	0.62	0.63	0.05	0.16	0.64	0.66
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(1)}$	0.00	0.03	0.62	0.63	0.01	0.04	0.67	0.73
	$\boldsymbol{\tau}^{(4)}, \Sigma^{(2)}$	0.01	0.04	0.62	0.62	0.01	0.04	0.63	0.63

In each of 24 simulation settings, we simulate 10,000 data sets under no unmeasured confounding with $I = 250$ pairs for the three outcome variables of interest and again use Huber's M-statistic. For each of the 8 combinations of treatment effects and covariances, closed testing is used to test individual hypotheses, and tests are run at $\Gamma = 1.25, 1.375$, and 1.5. We also include the power for rejecting the overall null for each combination and at each level of Γ. The values for the treatment effect vector and the covariances were as follows:

1. $\boldsymbol{\tau}^{(1)} = [0.2, 0.225, 0.25]$; $\boldsymbol{\tau}^{(2)} = [0.25, 0.3, 0.35]$; $\boldsymbol{\tau}^{(3)} = [0.2, 0.25, 0.35]$; $\boldsymbol{\tau}^{(4)} = [0.15, 0.25, 0.35]$

2. $\Sigma^{(1)} = \text{Diag}(1)$; $\Sigma^{(2)}_{ij} = 1$ if $i = j$, $\Sigma^{(2)}_{ij} = 0.5$ otherwise.

Table 13 shows the power for rejecting Fisher's sharp null for each outcome under four different vectors of true treatment effect values and two different forms of the covariance matrix. The magnitude of the improvement attained through the minimax procedure can be seen to depend on many factors. All else equal, as Γ increases the gains in power also increase. The gains in power tend to be more substantial in the *iid* cases ($\Sigma^{(1)}$) versus the positively correlated case ($\Sigma^{(2)}$), as for each intersection hypothesis the minimax procedure tends to resemble more closely the individual testing approach when there is positive correlation since the worst-case confounders across outcomes tend to align more closely. For example, with $\boldsymbol{\tau}^{(2)} = [0.25, 0.3, 0.35]$ at $\Gamma = 1.5$, the power after combining individual sensitivity analyses and after using the minimax procedure are $[0.16, 0.38, 0.64]$ versus $[0.22, 0.48, 0.76]$ when the paired differences are independent across outcomes, yet were $[0.18, 0.38, 0.64]$ versus $[0.20, 0.42, 0.68]$ when positively correlated. Gains are also most apparent when the treatment effects are of roughly the same magnitude ($\boldsymbol{\tau}^{(1)}$ and $\boldsymbol{\tau}^{(2)}$), while the gains tail off as one outcome increasingly determines the rejection of the overall null (compare $\boldsymbol{\tau}^{(2)}, \boldsymbol{\tau}^{(3)}, \boldsymbol{\tau}^{(4)}$). Thus, while the gains for testing the overall null hypothesis may be most apparent, the minimax procedure can provide meaningful improvements for testing nulls on individual outcomes.

In Appendix C.3, we show that our procedure does provide strong familywise error control in the presence of true intersection nulls as desired.

4.7. Improved Robustness to Unmeasured Confounding for Elevated Napthalene in Smokers

4.7.1. Conflicting Desires for the Worst-Case Confounder

To make concrete the factors allowing for the gains discussed in this work, Table 8 show the values and aligned ranks for $\log_e$ urinary concentrations of 1-naphthol and 2-naphthol for two individuals, one smoker and one nonsmoker, who were matched as a pair by the full match described in Appendix C.1. Both individuals are Hispanic males aged over 50, are similar in terms of height and weight, and are both exposed to PAHs occupationally, yet the

Table 8: Worst-Case Confounders in a Particular Pair at $\Gamma = 10$

	1-Naphthol				2-Naphthol			
	R_{ij1}	q_{ij1}	u^*_{ij1}	$\mathbb{E}[T_{i1}]$	R_{ij2}	q_{ij2}	u^*_{ij2}	$\mathbb{E}[T_{i2}]$
NS	6.39	353	0	1274	8.63	1350	1	1260
S	8.54	1366	1		7.07	363	0	
	Minimax							
			$\mathbf{u}^*$	$\mathbb{E}[T_{i1}]$	$\mathbb{E}[T_{i2}]$			
			$[0.562, 0]$	571	1137			

smoker (labeled S) has higher levels of 1-naphthol and lower levels of 2-naphthol.

The tests of both 1-naphthol and 2-naphthol had observed test statistics that were larger than their expectations under Fisher's sharp null with $\Gamma = 1$. Hence, the individual sensitivity analyses will choose the binary vector of $\mathbf{u}^*_k$ such that the individual with the larger observed response is given the value 1, thus having the higher probability of smoking. For 1-naphthol this is the smoker, but for 2-naphthol this is the nonsmoker, as is shown in Table 8. Although we do not know the value of this unmeasured confounder, we do know that logically, the unmeasured confounder cannot simultaneously increase the odds that individual 1 smokes relative to individual 2 and the odds that individual 2 smokes relative to individual 1. Simply combining these two sensitivity analyses would ignore the contradictory values of $\mathbf{u}^*_k$. Table 8 also gives the expectation of the test statistic for the individual outcomes assessed separately at $\Gamma = 10$, a value of Γ for which the minimax procedure rejects the overall null, but using Holm-Bonferroni to combine sensitivity analyses fails to reject. Conducting sensitivity analyses separately and allowing for an illogical effect of the unmeasured confounder, the worst-case expectations for the contribution from this matched set to the test statistics' expectations are 1274 and 1260 for 1- and 2-naphthol.

Recognizing that the unmeasured confounders must have the same impact on odds of treatment for individuals in a matched set yields markedly different results for the overall sensitivity analysis in this pair, as is demonstrated in the section labeled "Minimax" in Table 8. First, we note that the values of the unmeasured confounder for both individuals are fractional, an occurrence which is provably impossible when conducting sensitivity analyses

for any given outcome (Rosenbaum and Krieger, 1990). This makes the probabilities of assignment to treatment and control much less extreme than they possibly could have been: conditional on one of the two individuals receiving the treatment, the smoker is given a probability of $\exp\{\log(10)0.391\}/(\exp\{\log(10)0.391\} + \exp\{\log(10).953\}\}) = 0.22$ of being a smoker, while at $\Gamma = 10$ this probability could have been as low as $1/(1+10)$ and as high as $10/(1+10)$. In minimizing the maximal deviate, the optimization problem determined that a compromise should be made between the two conflicting desires of the individual level sensitivity analysis, but that it should favor making 2-naphthol more significant. Hence, we see that the contribution to the overall expectation of the two test statistics is larger than what it would have been at no unmeasured confounding for 2-naphthol (1137 vs 856.5), but is actually smaller for 1-naphthol (571 vs 859.5).

4.7.2. Sensitivity of Overall and Outcome Specific Effects

As was stated in Section 4.1.2, the conclusions of either of the individual level tests on 1- and 2-naphthol were both overturned at $\Gamma = 7.78$ when using Holm-Bonferroni. This is also the maximal level of Γ at which we can claim overall significance of at least one of these metabolites. The minimax procedure for testing the overall null hypothesis was able to claim robustness of this same finding up until $\Gamma = 10.22$, representing a substantial increase in robustness. In this application the overall null is of interest, as both naphthalene metabolites are indicators of naphthalene exposure. Hence, rejecting the overall null implies that we can suggest that at least one of our indicators of naphthalene exposure is significantly elevated for smokers relative to nonsmokers, even if we are not able to identify a particular metabolite that is significant at that level of unmeasured confounding.

To exploit the potential gains in power for individual tests of 1-naphthol and 2-naphthol, we use a closed testing procedure. In our example, doing so means that if we reject the null $H_1 \wedge H_2$ at level 0.05 through our minimax procedure we can then test the individual hypotheses H_1 and H_2 at level 0.05 (rather than 0.025) and still maintain the proper familywise error rate. Since our test of the overall null rejects until $\Gamma = 10.22$, the closed testing procedure

allows us to perform individual tests up to that level of unmeasured confounding. The individual tests of 1- and 2-naphthol *without* a Bonferroni correction (i.e., tested at $\alpha = 0.05$) were not overturned until a Γ of 7.83 and 8.20 respectively. As our minimax procedure rejects the overall null $H_1 \wedge H_2$ for all Γ between 7.78 and 8.20, we can declare improved robustness of the individual level tests. That is, we can reject the null of no effect for 1- and 2-naphthol at all levels of Γ up to $\Gamma = 7.83$ and 8.20, rather than $\Gamma = 7.78$.

4.8. Discussion

In a randomized clinical trial, counfounders not accounted for in the trial's design are, on average, balanced through randomized assignment of the intervention. As such, there is less of a concern that the observed results are driven by a causal mechanism other than the one under investigation. In observational studies, there is no such guarantee of balance on the unmeasured confounders between the two groups under comparison. When testing for a causal effect on multiple outcome variables, concerns about a loss of power by controlling the familywise error rate both under the assumption of no unmeasured confounding and within the sensitivity analysis may arise. We have demonstrated through this work that when dealing with multiple comparisons in a sensitivity analysis, the loss in power from controlling the familywise error rate can be attenuated.

As mentioned in Section 4.5, our method can be used in conjunction with sequential rejection procedures which proceed by rejecting intersection null hypotheses on a sequence of subsets of outcomes, $\{\mathcal{K}_\ell\}$. For certain types of null hypotheses, such as those for the value of an additive treatment effect with one sided alternatives, our method could also be used while employing the partitioning principle of familywise error control (Finner and Strassburger, 2002). One deficiency of our method is that it does not account for correlation between test statistics, which can greatly improve power in the presence of dependence (Westfall and Young, 1993; Romano and Wolf, 2005). While the simulation studies of Section 4.6 reveal marked improvements when test statistics are independent, these gains are far more modest when the test statistics are correlated and further improvements are desired. Deriving

methods for sensitivity analyses which exploit correlation between test statistics remains a topic of ongoing research. Another limitation is that our method can only be used for sensitivity analyses after matching, as the structure of matched sets returned by matching algorithms allows for a straightforward relationship between the assignment probabilities and the variances of our test statistics. In unmatched or stratified analyses, while the logical inconsistencies noted herein are still present, optimizing over the unknown assignment probabilities can no longer be expressed as a quadratically constrained linear program.

In our motivating example, we argue that if smoking causes increased naphthalene exposure, it would elevate levels of both 1- and 2-naphthol in the body. Though related, these metabolites are not affected equally by measured and unmeasured confounding variables: for example, there are certain genetic variants that are only believed to affect the prevalence of particular naphthalene metabolites (Yang et al., 1999). When focusing on a single outcome variable, the worst-case confounder is allowed to optimally align itself with the responses in each matched set through selecting the worst-case allocation of treatment assignment probabilities. If we are instead trying to disprove the overall truth of null hypotheses on multiple outcomes, the worst-case confounder likely cannot affect the treatment assignment probabilities in a way that simultaneously yields the worst-case inference for all outcomes. Exploiting this fact not only lends higher power to a sensitivity analysis for the overall null across all outcomes, but also increases power for testing hypotheses on individual outcomes through the use of certain sequential rejection procedures.

CHAPTER 5 : Sensitivity Analysis for the Average Treatment Effect in Matched Observational Studies

Inspired by work of Paul Rosenbaum

5.1. Introduction

In the analysis of observational studies, unease is sometimes expressed with the assumption of a constant treatment effect for each individual in the study; see, for example, Heckman et al. (2006) and Rosenbaum (2002c, Discussion and Rejoinder). To address this unease, we present a new method for conducting a sensitivity analysis for the *average treatment effect* in a paired observational study while allowing for heterogeneous individual effects. Through this work we hope to further facilitate the conducting of sensitivity analyses in the analysis of observational data, as in many fields the average treatment effect represents the most common quantification of intervention's impact (Imbens, 2004).

5.2. A Paired Observational Study

5.2.1. Notation for Paired Experiments and Observational Studies

There are I independent matched pairs. In each of i matched pairs, there is one individual who receives the treatment, denoted as $Z_{ij} = 1$, and one who receives the control, denoted as $Z_{ij} = 0$, such that $Z_{i1} + Z_{i2} = 1$ for each i. These matched pairs are formed on the basis of observed pre-treatment covariates $\mathbf{x}_{ij}$, so that $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each pair i; however, individuals may differ on the basis of an unobserved covariate u_{ij}, such that $u_{i1} \neq u_{i2}$. Each individual has a potential outcome under treatment, r_{Tij}, and under control, r_{Cij}. The fundamental problem of causal inference is that (r_{Tij}, r_{Cij}) are not jointly observable; rather, we observe the response $R_{ij} = r_{Tij}Z_{ij} + r_{Cij}(1 - Z_{ij})$ for each individual. See Neyman (1923) and Rubin (1974) for more on the potential outcomes framework.

Let Ω_I be the set of 2^I possible values of $\mathbf{Z}$ under the matched pairs design. In a paired

randomized experiment, randomness is modeled through the assignment vector; each $\mathbf{z} \in \Omega_I$ has probability 2^{-I} of being selected. Hence, quantities dependent on the assignment vector such as $\mathbf{Z}$, $\mathbf{R}$ are random, whereas $\mathcal{F}_I = \{(r_{Tij}, r_{Cij}, x_{ij}, u_{ij}\}$ contains fixed quantities.

For a randomized experiment, $\pi_i := \mathbb{P}(Z_{i1} = 1 | \mathcal{F}_I, \mathbf{Z} \in \Omega_I) = 1/2$. In an observational study, it may be the case that $\pi_i \neq 1/2$ due to differences in u_{i1} and u_{i2} for the individuals in matched pair i of I. As such, the probability of an observed allocation $\mathbf{z}$ must instead be written as:

$$\mathbb{P}(\mathbf{Z} = \mathbf{z} | \mathcal{F}_I, Z \in \Omega_I) = \prod_{i=1}^{I} \pi_i^{z_{i1}} (1 - \pi_i)^{1 - z_{i1}}$$

Without control over the assignment mechanism, the probabilities π_i are unknown to the researcher. Through a *sensitivity analysis*, one seeks to assess the robustness of a study's finding to departures from an idealized paired experiment. A sensitivity analysis places bounds on the allowable departure from a pure randomized experiment for two individuals in the same matched pair. We use the model of Rosenbaum (1987), which controls the allowable departure from a paired randomized experiment through a parameter $\Gamma \geq 1$. In each matched pair, we bound π_i above and below by

$$\frac{1}{1 + \Gamma} \leq \pi_i \leq \frac{\Gamma}{1 + \Gamma}.$$

This model can be derived as a simplification of the model in Rosenbaum (2002a, Section 4) in the case of matched pairs. The sensitivity analysis proceeds by, for a given value of Γ, finding the worst-case null distribution for the inferential problem at hand. One then iteratively increases the value of Γ until the null hypothesis can no loner be rejected. This changepoint Γ then serves as a measure of robustness of the study's findings to unmeasured confounding.

5.3. The Average Treatment Effect

Define $\varphi_{i1} = r_{Ti1} - r_{Ci2}$ and $\varphi_{i2} = r_{Ti2} - r_{Ci1}$ to be the observed paired difference if the $Z_{i1} = 1$ and $Z_{i1} = 0$ respectively. The treated minus control difference in pair i that is actually observed can then be written as:

$$Y_i = Z_{i1}(\varphi_{i1}) + (1 - Z_{i1})\varphi_{i2}$$

The *average treatment effect* in a paired experiment or observational study is defined as

$$\begin{aligned}\bar{\Delta} &:= \frac{1}{2I}\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{Tij} - r_{Cij}) \\ &= \frac{1}{2I}\sum_{i=1}^{I}(\varphi_{i1} + \varphi_{i2})\end{aligned}$$

We consider the estimator $\bar{Y} = I^{-1}\sum_{i=1}^{I} Y_i$, and use it henceforth as a test statistic for inference on $\bar{\Delta}$. In a purely randomized matched pairs design with $\pi_i = 1/2$, $\bar{Y}$ is an unbiased estimator of $\bar{\Delta}$ (Rosenbaum, 2002a). In an observational study, $\pi_i = 1/2$ would represent a specious assumption and $\bar{Y}$ may well be biased for $\bar{\Delta}$.

If we were to further assume an additive treatment effect model, $r_{Tij} = r_{Cij} + \tau$, then a null hypothesis on $\bar{\Delta}$ would be *sharp*, in that it would entirely specify the pairs (r_{Tij}, r_{Cij}) for each individual, hence facilitating the use of randomization tests to assess statistical significance and allowing one to use the methods of Rosenbaum (2007) to conduct a sensitivity analysis for $\bar{Y}$. In the absence of an assumption of additivity, a null hypothesis $H_0 : \bar{\Delta} = \Delta_0$ is *composite*, in that there are in fact infinitely many allocations for the $2I$ missing potential outcomes that satisfy the null in question. We call a set of potential outcomes *consistent* with the null in question if the following three conditions hold.

(A1) Consistency with observed data: $Z_{i1}\varphi_{i1} + (1 - Z_{i1})\varphi_{i2} = Y_i$

(A2) Consistency with additional assumptions made on potential outcomes (for example, additive treatment effect; nonnegative treatment effect)

(A3) Agreement with the null hypothesis: $\sum_{i=1}^{I}(\varphi_{i1} + \varphi_{i2}) = 2I\bar{\Delta}_0$

The first condition recognizes that we know the true values for half of the potential outcomes based on the observed data. The second condition means that if the practitioner has made additional assumptions on the potential outcomes, those assumptions must be satisfied in the allocations of potential outcomes under consideration. The third condition signifies that when testing a null hypothesis, we must only consider allocations of potential outcomes where the corresponding causal parameter takes on the desired value.

Let $\mathcal{H}(\bar{\Delta}_0)$ represent the set of potential outcomes satisfying conditions A1 - A3. As the size of a composite null hypothesis test is the supremum of the sizes of the elements of the composite null, to reject the null $H_0 : \bar{\Delta}_0$ at level α, we must reject the null for all $\{\varphi_{i1}, \varphi_{i2}\} \in \mathcal{H}(\bar{\Delta}_0)$ at level α. As will now be made clear, such a pursuit would be a fool's errand without a further restriction on the set of consistent allocations of potential outcomes over which we aspire towards type I error control.

Example 1 (Motivating a Further Restriction)**.** *Suppose without loss of generality* $\bar{y} > \bar{\Delta}_0$*. Let* m_i *be the missing paired difference in matched set* i*, and set* m_i *as*

$$m_i = \begin{cases} (2\bar{\Delta}_0 - \bar{y}) + I\max\{|y_i|\} & i \in \{1, ..., I/2\} \\ (2\bar{\Delta}_0 - \bar{y}) - I\max\{|y_i|\} & i \in \{I/2+1, ..., I\} \end{cases}$$

Clearly, $\{\varphi_{i1}, \varphi_{i2}\} \in \mathcal{H}(\bar{\Delta}_0)$*. However, for this allocation, under no unmeasured confounding we see that* $\mathbb{P}(\bar{Y} \geq \bar{y}|\mathcal{F}_I, \mathbf{Z} \in \Omega_I) > \mathbb{P}\left(\sum_{i=1}^{I/2} Z_{i1} > \sum_{i=I/2+1}^{I} Z_{i1}|\mathcal{F}_I, \mathbf{Z} \in \Omega_I\right) \to 0.5$ *as* $I \to \infty$*. Furthermore, this probability could be made strictly larger in the corresponding sensitivity analysis.*

This problem plagues not only the analysis of observational studies, but even the analysis

of randomized experiments through the potential outcomes framework. In contemplating how to proceed, we now discuss, and subsequently borrow from, the standard procedure for inference on the average treatment effect in randomized experiments.

5.3.1. Asymptotic Normality and Estimating an Upper Bound on the Variance

In this section, we describe the subset of the composite null $\mathcal{H}_0(\bar{\Delta}_0)$ over which we will perform both inference under no unmeasured confounding and a sensitivity analysis. Our first condition will be fairly benign, while the second will deserve closer consideration. We initially restrict attention to elements of the composite null for which the estimator $\bar{Y}$ is asymptotically normal. One set of conditions given by Hájek and Sǐdák (1967) is as follows.

Proposition 1 (Hájek and Sǐdák (1967))**.** *If* $\sum_{i=1}^{I}(\varphi_{i1}-\varphi_{i2})^2 / \max_{1\leq i\leq I}(\varphi_{i1}-\varphi_{i2})^2 \to \infty$ *and* $I^{-1}\sum_{i=1}^{I}(\varphi_{i1}-\varphi_{i2})^2 \to \eta > 0$ *then*
$I\bar{Y} \xrightarrow{d} \mathcal{N}\left(\sum_{i=1}^{I}(\pi_i\varphi_{i1}+(1-\pi_i)\varphi_{i2}), \sum_{i=1}^{I}\pi_i(1-\pi_i)(\varphi_{i1}-\varphi_{i2})^2\right)$

Assuming these necessary conditions for asymptotic normality does not alleviate the problems raised in the previous section, as the allocation of missing potential outcomes given in Example 1 satisfies these conditions. Nonetheless, employing asymptotic normality when conducting inference does lead to a natural additional condition to impose to make the problem tractable. Under a normal approximation and assuming no unmeasured confounding, the distribution of $\bar{Y}$ generated by different elements $\mathcal{H}(\bar{\Delta}_0)$ only differs due to their effect on the variance of $\bar{Y}$. When using the potential outcomes framework in randomized experiments, randomization inference for the average treatment effect typically proceeds by finding a consistent estimator of an upper bound on the variance of the estimated ATE, and using that variance to conduct inference under a normal approximation; see Neyman (1923) and Ding (2014) among many. Might a similar approach be employed in the analysis of the average treatment effect in observational studies?

Under a matched pairs design, the variance of the average treatment effect is given by:

$$\text{var}(\bar{Y}|\mathcal{F}_I, \mathbf{Z} \in \Omega_I) = I^{-2} \sum_{i=1}^{I} \pi_i(1-\pi_i)(\varphi_{i1} - \varphi_{i2})^2$$

This variance is unknown not only due to the unobserved potential outcomes, but also due to the fact that the in an observational study, the values of $\{\pi_i\}$ are unknown to the researcher. Nonetheless, we now demonstrate that for a given maximal allowable departure from a matched pairs design Γ, one can similarly upper bound the variance in a sensitivity analysis.

Suppose we are testing the null that $\bar{\Delta} = \bar{\Delta}_0$, and consider the estimator $\bar{\bar{V}}_\Gamma := 2\Gamma/(1+\Gamma)\sum_{i=1}^{I}(Z_{i1}(\varphi_{i1} - \bar{\Delta}_0)^2 + (1-Z_{i1})(\varphi_{i2} - \bar{\Delta}_0)^2)$.

Proposition 2.

$$\mathbb{E}[\bar{\bar{V}}_\Gamma|\mathcal{F}_I, \mathbf{Z} \in \Omega_I] \geq I^2 var(\bar{Y}|\mathcal{F}_I, \mathbf{Z} \in \Omega_I)$$

Proof.

$$\begin{aligned}
\mathbb{E}[\bar{\bar{V}}_\Gamma|\mathcal{F}_I, \mathbf{Z} \in \Omega_I] &= 2\Gamma/(1+\Gamma)\sum_{i=1}^{I}(\pi_i(\varphi_{i1} - \bar{\Delta}_0)^2 + (1-\pi_i)(\varphi_{i2} - \bar{\Delta}_0)^2) \\
&\geq 2\sum_{i=1}^{I}\pi_i(1-\pi_i)((\varphi_{i1} - \bar{\Delta}_0)^2 + (\varphi_{i2} - \bar{\Delta}_0)^2) \\
&\geq \sum_{i=1}^{I}\pi_i(1-\pi_i)((\varphi_{i1} - \bar{\Delta}_0)^2 + (\varphi_{i2} - \bar{\Delta}_0)^2 - 2(\varphi_{i1} - \bar{\Delta}_0)(\varphi_{i2} - \bar{\Delta}_0) \\
&= \sum_{i=1}^{I}\pi_i(1-\pi_i)(\varphi_{i1} - \varphi_{i2})^2 = I^2\text{var}(\bar{Y}|\mathcal{F}_I, \mathbf{Z} \in \Omega_I)
\end{aligned}$$

□

Let $\bar{\bar{v}}_\Gamma$ denote the observed value of the random variable $\bar{\bar{V}}_\Gamma$. Moving forward, we proceed with inference for the composite null containing potential outcomes such that the following three conditions hold:

1. $\{\varphi_{i1}, \varphi_{i2}\} \in \mathcal{H}_0(\bar{\Delta}_0)$

2. $\sum_{i=1}^{I} \pi_i(1-\pi_i)(\varphi_{i1}-\varphi_{i2})^2 \leq \bar{\bar{v}}_\Gamma$

3. $\bar{Y}$ is asymptotically normal.

5.4. Sensitivity Analysis for the Average Treatment Effect

Without loss of generality, suppose the our estimate of the average treatment effect exceeds its null expectation at $\Gamma = 1$, i.e. $\bar{y} > \bar{\Delta}_0$. Further, assume for notational convenience that in each pair the first individual received the treatment so $Z_{i1} = 1 \ \ \forall i$. Hence, φ_{i1} is known and φ_{i2} is unknown.

Employing a normal approximation for the average treatment effect, consider the following optimization problem

$$
\begin{aligned}
\underset{\{\pi_i, \varphi_{i2}\}}{\text{minimize}} \ & \frac{\sum_{i=1}^{I} \varphi_{i1} - \sum_{i=1}^{I} \left(\pi_i \varphi_{i1} + (1-\pi_i)\varphi_{i2}\right)}{\sqrt{\sum_{i=1}^{I} \pi_i(1-\pi_i)(\varphi_{i1}-\varphi_{i2})^2}} \\
\text{subject to} \ & \sum_{i=1}^{I} \varphi_{i1} + \varphi_{i2} = 2I\bar{\Delta}_0 \\
& \sum_{i=1}^{I} \pi_i(1-\pi_i)(\varphi_{i1}-\varphi_{i2})^2 \leq \bar{\bar{v}}_\Gamma \\
& \frac{1}{1+\Gamma} \leq \pi_i \leq \frac{\Gamma}{1+\Gamma}
\end{aligned}
\tag{P1}
$$

Problem (P1) encodes the desired sensitivity analysis at level of unmeasured confounding Γ, as under the normal approximation minimizing the standardized deviate is equivalent to maximizing the p-value for the performed hypothesis test. Unfortunately, the above problem is not convex. We will now take steps to facilitate its computation.

We begin by, through the following lemma, simplifying the optimization problem with respect to the unknown $\{\pi_i\}$

Lemma 1. *Suppose* $1/(1+\Gamma) \leq \pi_i \leq \Gamma/(1+\Gamma)\ \forall i$. *Then,* $\mathbb{P}(\bar{Y} \geq \bar{y}) \leq \mathbb{P}(I^{-1}\sum_{i=1}^{I} \tilde{Y}_i \geq \bar{y})$, *where* $\tilde{Y} = \tilde{Z}_i \max\{\varphi_{i2}, \varphi_{i1}\} + (1-\tilde{Z}_i)\min\{\varphi_{i2}, \varphi_{i1}\}$, *and* $\tilde{Z}_i \overset{iid}{\sim} Bern(\Gamma/(1+\Gamma))$.

Proof. For any $\pi_i \in [1/(1+\Gamma), \Gamma/(1+\Gamma)]$, $\mathbb{P}(Y_i = \max\{\varphi_{i2}, \varphi_{i1}\}) \leq \mathbb{P}(\tilde{Y}_i = \max\{\varphi_{i2}, \varphi_{i1}\})$. As Y_i and $\tilde{Y}_i$ only take on the values $\min\{\varphi_{i2}, \varphi_{i1}\}$ and $\max\{\varphi_{i2}, \varphi_{i1}\}$, $\tilde{Y}_i$ is stochastically larger than Y_i. The result then follows from preservation of stochastic ordering under independent convolutions. □

For any fixed values of $\{\varphi_{i1}, \varphi_{i2}\}$, the worst-case unmeasured confounder would thus attribute $\pi_i = \frac{\Gamma}{1+\Gamma}$ if $\varphi_{i1} \geq \varphi_{i2}$, and $\pi_i = \frac{1}{1+\Gamma}$ otherwise. This suggests that instead of optimizing over $\pi_i \in [1/(1+\Gamma), \Gamma/(1+\Gamma)]$ we can express π_i as a function of φ_{i1} and φ_{i2}, $\pi_i(\varphi_{i1}, \varphi_{i2}) = w_i/(1+\Gamma) + (1-w_i)\Gamma/(1+\Gamma)$, where $w_i = \mathbb{1}\{\varphi_{i2} \geq \varphi_{i1}\}$. Before proceeding as such, we need to ensure that for any allocation of potential outcomes and true treatment assignment probabilities satisfying the constraints of Problem (P1), the corresponding worst-case allocation also satisfies the above constraints. This is indeed true, as the following trivial lemma indicates:

Lemma 2. *For any* $\{\pi_i\} \in [1/(1+\Gamma), \Gamma/(1+\Gamma)]^I$*:*

$$\sum_{i=1}^{I} \pi_i(1-\pi_i)(\varphi_{i1}-\varphi_{i2})^2 \geq \frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I}(\varphi_{i1}-\varphi_{i2})^2$$

Hence, for any allocation of $\{\pi_i, \varphi_{i1}, \varphi_{i2}\}$ with a variance for the estimated average treatment effect that is less than or equal to the variance upper bound, the worst-case distribution based upon $\{\varphi_{i1}, \varphi_{i2}\}$ has a variance that is also less than or equal to the variance upper bound.

Lemmas 1 and 2 allows us to consider the following simplified optimization problem:

$$
\begin{aligned}
\underset{\{\pi_i,\varphi_{i2}\}}{\text{minimize}} \quad & \frac{\sum_{i=1}^{I}\varphi_{i1} - \sum_{i=1}^{I}\left(\pi_i\varphi_{i1} + (1-\pi_i)\varphi_{i2}\right)}{\sqrt{\frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I}(\varphi_{i1}-\varphi_{i2})^2}} \\
\text{subject to} \quad & \sum_{i=1}^{I}\varphi_{i1} + \varphi_{i2} = 2I\bar{\Delta}_0 \\
& \frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I}(\varphi_{i1}-\varphi_{i2})^2 \leq \bar{\bar{v}}_\Gamma \\
& \pi_i = w_i/(1+\Gamma) + (1-w_i)\Gamma/(1+\Gamma) \\
& w_i = \mathbb{1}\{\varphi_{i2} \geq \varphi_{i1}\}
\end{aligned}
\tag{P2}
$$

The above problem could be formulated as an integer program through the use of a "Big-M" formulation, as is discussed in Section 5.5; however, such formulations have notoriously weak continuous relaxations and can thus be very slow in practice for even moderately sized problems (Bertsimas and Tsitsiklis, 1997). Fortunately, such an approach is not necessary. In fact, as we now demonstrate, a solution to problem (P1) can be attained in $\mathcal{O}(I)$ operations.

5.4.1. A Linear Time Algorithm

To proceed, define C_s^+ and C_s^-, $s \in \{1, ..., I-1\}$, as

$$
C_s^+ = \frac{2I\bar{\Delta}_0 - 2\sum_{i=1}^{I}\varphi_{i1} - C_s^-(I-s)}{s}
$$

$$
C_s^- = \frac{4\sum_{i=1}^{I}(\bar{\Delta}_0 - \varphi_{i1})\frac{I-s}{s} - 2\sqrt{\left(\frac{I-s}{s}\right)\left(I\frac{(1+\Gamma)^2}{\Gamma}\bar{\bar{v}}_\Gamma - 4\left(\sum_{i=1}^{I}(\bar{\Delta}_0 - \varphi_{i1})\right)^2\right)}}{2I\left(\frac{I-s}{s}\right)}
$$

Furthermore, define $\mu_0, \mu_1, ..., \mu_{I-1}$ and $\nu_0^2, \nu_1^2, ..., \nu_{I-1}^2$ as:

$$\mu_0 = \begin{cases} \frac{2}{1+\Gamma}\sum_{i=1}^{I}(\varphi_{i1} - \bar{\Delta}_0) & \frac{\Gamma}{(1+\Gamma)^2}4I^{-1}(\sum_{i=1}^{I}(\bar{\Delta}_0 - \varphi_{i1}))^2 \leq \bar{\bar{v}} \\ -\infty & \text{otherwise} \end{cases},$$

$$\nu_0^2 = \frac{\Gamma}{(1+\Gamma)^2}4I^{-1}(\sum_{i=1}^{I}(\bar{\Delta}_0 - \varphi_{i1}))^2$$

and, for $s \in \{1, ..., I-1\}$,

$$\mu_s = \sum_{i=1}^{I}\varphi_{i1} + \frac{\Gamma C_s^+}{1+\Gamma}s + \frac{C_s^-}{1+\Gamma}(I-s)$$
$$\nu_s^2 = \bar{\bar{v}}$$

Finally, define the deviate a as:

$$a := \min_{s\in\{0,...,I-1\}} \frac{I\bar{y} - \mu_s}{\nu_s}$$

Theorem 2. *Suppose that* $\sum_{i=1}^{I}(\varphi_{i1} - \varphi_{i2})^2 / \max_{1\leq i\leq I}(\varphi_{i1} - \varphi_{i2})^2 \to \infty$ *and* $I^{-1}\sum_{i=1}^{I}(\varphi_{i1} - \varphi_{i2})^2 \to \eta > 0$. *Consider conditional probabilities of receiving the treatment* $1/(1+\Gamma) \leq \pi_i \leq \Gamma/(1+\Gamma)$ *for each matched pair. Then, for any allocation of potential outcomes such that (a)* $\{\varphi_{i1}, \varphi_{i2}\} \in \mathcal{H}_0(\bar{\Delta}_0)$ *and (b)* $\sum_{i=1}^{I}\pi_i(1-\pi_i)(\varphi_{i1} - \varphi_{i2})^2 \leq \bar{\bar{v}}$*:*

$$\lim_{I\to\infty} \mathbb{P}(\bar{Y} \geq \bar{y} | \mathcal{F}_I, \mathbf{Z} \in \Omega_I) \leq 1 - \Phi(a),$$

where $\Phi(\cdot)$ *is the standard normal CDF. That is, the deviate* a *is the solution to Problem (P1).*

The proof is deferred to Appendix D.1. This algorithm is similar in spirit to the one presented in Rosenbaum (2002b) for conducting a sensitivity analysis for the attributable effect, in that a seemingly complicated optimization problem over a composite null hypothesis can be reduced to a small number of simple evaluations.

5.5. Known Direction of Effect

Oftentimes it is reasonable to assume that while a treatment may have heterogeneous effects from one individual to the next, the direction of the effect lies in the same direction for all individuals. Without loss of generality, we will proceed assuming the treatment effect is nonnegative for each individual, i.e. that $r_{Tij} \geq r_{Cij}\ \forall i, j$. This restriction can be added to Problem (P2) as follows:

$$
\begin{aligned}
\underset{\{\pi_i, \varphi_{i2}\}}{\text{minimize}}\ & \frac{\sum_{i=1}^{I} \varphi_{i1} - \sum_{i=1}^{I} \left(\pi_i \varphi_{i1} + (1-\pi_i)\varphi_{i2}\right)}{\sqrt{\frac{\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{I} (\varphi_{i1} - \varphi_{i2})^2}} \\
\text{subject to}\ & \sum_{i=1}^{I} \varphi_{i1} + \varphi_{i2} = 2I\bar{\Delta}_0 \\
& \frac{\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{I} (\varphi_{i1} - \varphi_{i2})^2 = \bar{\bar{v}}_\Gamma \\
& \pi_i = w_i/(1+\Gamma) + (1-w_i)\Gamma/(1+\Gamma) \\
& w_i = \mathbb{1}\{\varphi_{i2} \geq \varphi_{i1}\} \\
& \varphi_{i2} \geq -\varphi_{i1}
\end{aligned}
\tag{P3}
$$

An area of ongoing research is to seek a computationally scalable manner for solving Problem (P3). We can find the worst-case *expectation* for the estimated average treatment effect by solving the following integer program.

$$
\begin{aligned}
\underset{\{w_i, x_i^+, x_i^-\}}{\text{maximize}} \quad & \frac{\Gamma}{1+\Gamma}\sum_{i=1}^{I}\varphi_{i1} + \sum_{i=1}^{I}\left(\frac{1}{1+\Gamma}x_i^- + \frac{\Gamma}{1+\Gamma}x_i^+ - \frac{\Gamma-1}{1+\Gamma}\varphi_{i1}w_i\right) \qquad \text{(P4)} \\
\text{subject to} \quad & \sum_{i=1}^{I}\varphi_{i1} + x_i^- + x_i^+ = 2I\bar{\Delta}_0 \\
& -(1-w_i)\varphi_{i1} \leq x_i^- \leq (1-w_i)\varphi_{i1} \\
& w_i\varphi_{i1} \leq x_i^+ \leq w_i(2I\bar{\Delta}_0 - \varphi_{i1}) \\
& \frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I}(\varphi_{i1} - x_i^- - x_i^+)^2 \leq \bar{\bar{v}}_\Gamma \\
& w_i \in \{0,1\} \;\; \forall i \\
& x_i^+ + x_i^- \geq -\varphi_{i1} \;\; \forall i
\end{aligned}
$$

A conservative approach to finding the worst-case deviate would then be to simply use the variance upper bound to create, denoting the solution of Problem (P4) by $\bar{\bar{\mu}}$

$$a = \frac{I\bar{y} - \bar{\bar{\mu}}}{\sqrt{\bar{\bar{v}}_\Gamma}}.$$

One instance in which this is provably conservative is under the null $\bar{\Delta} = 0$. In the case of a nonnegative treatment effect, the only allocation of potential outcomes satisfying this null is that under Fisher's sharp null of no effect. Hence, the standard sensitivity analysis for Fisher's sharp null then yields a valid sensitivity analysis for the null of $\bar{\Delta} = 0$.

5.6. Bigger Effect for Individuals More Likely to Receive Treatment

One particular case of heterogeneity often considered in economics as an argument against assuming an additive treatment effect is known as "essential heterogeneity," wherein individuals who will benefit more from a given treatment are more likely to decide to take said treatment (Heckman et al., 2006). In the context of a paired observational study, this restriction could be written in the form $\pi_i - (1-\pi_i) \geq 0 \Leftrightarrow (r_{Ti1} - r_{Ci1}) \geq (r_{Ti2} - r_{Ci2})$. One

might think that such a consideration would impose a further constraint on a sensitivity analysis; however, the following proposition demonstrates, this turns out not to be the case.

Proposition 3. *The solutions to Problem (P1) under no assumption on the direction of effect and to Problem (P3) under a known direction of effects are also solutions under the constraint that the individual in a matched pair with the higher treatment effect has the higher probability of receiving the treatment*

Proof. We begin without assuming a known direction of effect. The solution returns worst-case $\varphi_{i2} = r_{Ti2} - r_{Ci1}$. The values of r_{Ti1} and r_{Ci2} are fixed. Suppose $\pi_i = \Gamma/(1+\Gamma)$ (the proof for $\pi_i = 1/(1+\Gamma)$ is analogous). To make it the case that $r_{Ti1} - r_{Ci1} \geq r_{Ti2} - r_{Ci2}$, set $r_{Ti2} = c + \varphi_{i2}$, set $r_{Ci1} = c$, and simply solve for the c such that the two treatment effects are equal. Doing so, we have $r_{Ti1} - c = c + \varphi_{i2} - r_{Ci2} \Rightarrow c = (r_{Ti1} + r_{Ci2} - \varphi_{i2})/2$. For any $c' < c$, $r_{Ti1} - r_{Ci1} \geq r_{Ti2} - r_{Ci2}$ as desired.

Under the assumption of non-negativity, we know that for any solution to Problem (P3) $\varphi_{i1} + \varphi_{i2} \geq 0$. Equivalently, this implies that $r_{Ti1} - r_{Ci1} + r_{Ti2} - r_{Ci2} \geq 0$. Suppose that $\pi_i = \Gamma/(1+\Gamma)$ (the proof for $\pi_i = 1/(1+\Gamma)$ is analogous). Then, setting $r_{Ti2} = r_{Ci2}$ and $r_{Ti1} - r_{Ci1} = \varphi_{i1} + \varphi_{i2}$ satisfies the constraint imposed by essential heterogeneity.

□

Hence, we can interpret the methods for a sensitivity analysis developed in the previous section as encompassing this particular form of heterogeneity.

5.7. Simulation: The Impact of Assumptions on Sensitivity to Unmeasured Confounding

In this section, we assess the power of a sensitivity analysis for the average treatment effect under the assumptions of additivity, nonnegative treatment effects, and no known direction of effect. We borrow the simulation setting of Rosenbaum (2005), which sought to assess the role of heterogeneity reduction in reducing sensitivity to unmeasured confounding under

an additive treatment effect model. In this simulation study, we hope to not only compare sensitivity analyses for the average treatment effects under assumptions of varying strength on the potential outcomes, but also to assess the role of heterogeneity reduction in reducing sensitivity to unmeasured confounding when an additive treatment effect model is not assumed.

In evaluating these assumptions, our simulation study assumes as is advocated in Rosenbaum (2004, 2007) that unbeknownst to the practitioner the paired data at hand truly arose from a paired randomized experiment (i.e., $\Gamma = 1$). The practitioner, blind to this, would like to not only perform inference under the assumption of no unmeasured confounding, but also assess the robustness of the study's findings to unmeasured confounding.

In the first simulation setting, called *larger, more heterogeneous* (**LM**), we draw, in each iteration, $I = 400$ paired differences Y_i with $Y_i \overset{iid}{\sim} \mathcal{N}(1/2, 1)$. In the second, called *smaller, less heterogeneous* (SL), we draw $I = 100$ paired differences with $Y_i \overset{iid}{\sim} \mathcal{N}(1/2, (1/2)^2)$. In both LM and SL, the estimated average treatment effect $\bar{Y}$, is distributed as $\bar{Y} \sim \mathcal{N}(1/2, 1/400)$; the settings differ only in the heterogeneity of the observed paired differences.

In each setting, we simulate 1000 data sets. We then perform a sensitivity analysis for a range of Γ, and assess the probability of correctly rejecting the null hypothesis of (a) $\bar{\Delta}_0 = 0$ and (b) $\bar{\Delta}_0 = 0.1$.

The results are shown in Figure 4. We first compare within the LM and SL settings. We note that for both null hypotheses in question, the sensitivity analysis for the average treatment effect without assumptions on the potential outcomes is less powerful than that performed under both an additive treatment effect model and a nonnegative treatment effect model. For the null $\bar{\Delta}_0 = 0$, we see that the additive treatment effect model and the nonnegative treatment effect model coincide, as here the null of zero average treatment effect under a nonnegative treatment effect model implies that Fisher's sharp null holds. When the null in question is $\bar{\Delta}_0 = 0.1$, we see that the sensitivity analysis assuming an additive treatment

effect model is more powerful than that assuming a nonnegative treatment effect, which is in turn more powerful than that without an assumption of a known direction of effect.

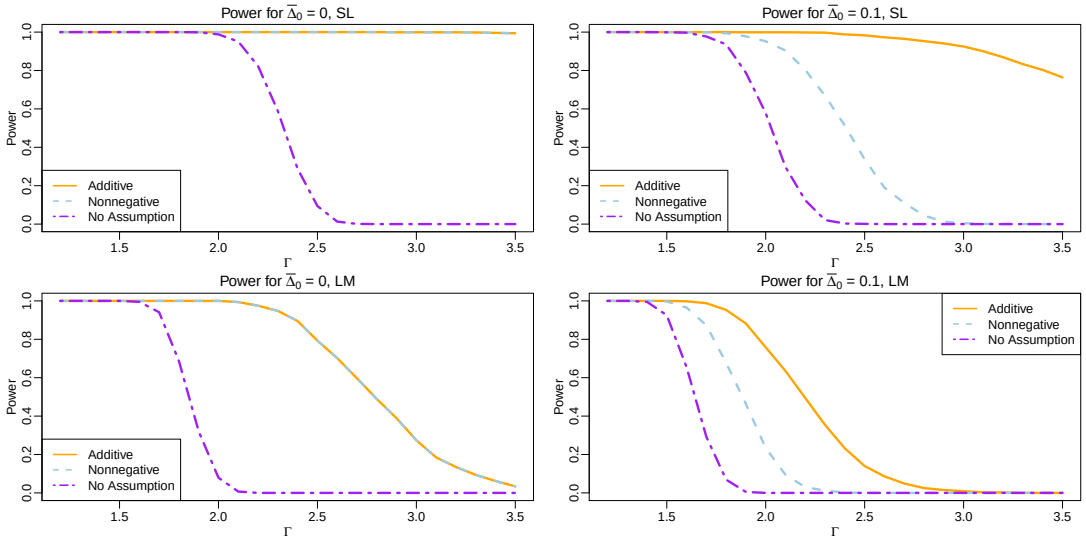

Figure 4: Power of a Sensitivity Analysis for the Average Treatment Effect. This figure shows the power for a sensitivity analysis as a function of Γ for testing null hypotheses on average treatment effect under additivity, a nonnegative treatment effect, and without further assumptions under scenarios LM and SL.

Now looking between the LM and SL settings, we note that setting SL yields a more powerful sensitivity analysis than LM for all three sets of assumptions on the potential outcomes; however, it appears as though the differences are most drastic under the assumption of additivity, and less so under the other two assumptions. Nonetheless, this settings lends further support to the importance of matching in observational studies as a means of reducing heterogeneity.

5.8. Discussion

In this work, we develop methods for conducting a sensitivity analysis for the average treatment effect with and without assuming a known direction of effect. This work indicates that not assuming additivity weakens the power of a sensitivity analysis to unmeasured confounding, which is consistent with the findings in Chapter 3 of this dissertation with respect to sensitivity analyses for the risk difference. These results should, by no means, be viewed as an indictment of additivity as a useful model for treatment effects. For example, even if

the overall assumption of additivity were to fail due to the presence of effect modification, it may be plausible for individuals within a given subgroup defined by the effect modifiers. In this case, Hsu et al. (2013, 2015) present methods for both discovering effect modification and subsequently testing for additivity within subgroups defined by the effect modifiers. Furthermore, Rosenbaum (2002c, Rejoinder, Section 3) notes that randomization inference assuming an additive treatment effect model is also the only non-parametric inference under the assumption that the marginal distributions of treatment and control potential outcomes adhere to an additive shift model for fixed values of the covariates. It is our belief that rather than supplanting the model of additivity, analyses with and without said assumption should be presented jointly as a means of further elucidating evidence for the strength of a given causal effect. As noted inRosenbaum (2002c, Rejoinder, Section 6), even if one is not certain that an additive treatment effect holds, confidence intervals for an additive treatment effect can nonetheless illustrate which additive effects are *not* plausible.

CHAPTER 6 : Discussion

One immediate area for future research is an extension of the work in Chapter 4 of the dissertation to methods for multiple comparisons which utilize sums of test statistics rather than the maxima of test statistics, and is joint work with Matt Olson and Dylan Small. Borrowing notation from Chapter 4 of this dissertation, and letting $\boldsymbol{\mu}(\boldsymbol{\varrho})$ and $\Sigma(\boldsymbol{\varrho})$ be the mean vector and covariance matrix for the test statistics, we seek to solve the following problem:

$$\begin{aligned}
&\min_{\varrho_{ij}, s_i} \ \max_{\lambda_k} \ \frac{\left(\boldsymbol{\lambda}^T(\mathbf{t} - \boldsymbol{\mu}(\boldsymbol{\varrho}))\right)^2}{\boldsymbol{\lambda}^T \Sigma(\boldsymbol{\varrho}) \boldsymbol{\lambda}} \\
&\text{subject to } \sum_{j=1}^{n_i} \varrho_{ij} = 1 \ \ \forall i \\
&\qquad\qquad \varrho_{ij} \geq 0 \ \ \forall i, j \\
&\qquad\qquad s_i \leq \varrho_{ij} \leq \Gamma s_i \ \ \forall i, j \\
&\qquad\qquad ||\boldsymbol{\lambda}|| \leq 1
\end{aligned}$$

Additional constraints on $\boldsymbol{\lambda}$ (say, $\lambda_k \geq 0$) can yield a one-sided test. We have demonstrated that, utilizing a projected subgradient descent algorithm, the above problem can be solved expeditiously. This procedure yields improved power in finite samples in the presence of strongly correlated outcome variables, and can be shown to, asymptotically, have higher design sensitivity for the test of the overall null than the procedure presented in Chapter 4.

More generally, in the next few years I hope to assess the extent to which clever applications of optimization routines can help quantify the relative merits of various approaches for the design and analysis of observational studies. Much advice on research designs and strategies for judging causality exists outside of statistics; however, as noted in Rosenbaum (2004), such advice is not always explicitly tied to tangible benefits for the resulting analysis. Rosenbaum (2004) discusses how exhibiting multiple operationalism and dose-response rela-

tionships yields enhanced robustness against unmeasured confounding as measured by both design sensitivities and the power of the resulting sensitivity analysis. Yet another example of a strategy yielding a quantifiable benefit is the use of "control" outcomes, i.e. outcomes known to be unaffected by the treatment (in the sense described in Rosenbaum (2002a, Section 6) that a purported significant treatment effect on the control outcome would make us question the study's design more strongly than our belief in the absence of an effect). In Rosenbaum (1992), it is demonstrated that through convex optimization control outcomes can be used to confidently eliminate certain types of unmeasured confounding. This has potential to strengthen the evidence in favor of a proposed causal mechanism, as the search for the worst-case unmeasured confounder is now limited to those which do not yield a significant result for the control outcome. This strategy can be made actionable through the formulation presented in Chapter 4, as it merely requires an additional quadratic constraint in the optimization problem. I hope to explore the extent to which known directions of effect and known directions of bias can also be exploited in this manner, hence furthering the connection between qualitative advice and quantitative improvement for causal inference in matched observational studies.

This thesis has investigated the role of modern optimization in the design and analysis of observational studies. Several of the methods presented herein require the solution of integer programs, which are $\mathcal{NP}$-hard in general. Owing to this, many of these proposed methods (and with them, certain chapters of this dissertation) may have been eschewed by statisticians on the grounds of practicability in the past. Through my work, I have come to the conclusion that these perceptions of old must be revisited and revised. In fact, over the past 25 years, a combination of algorithmic advances and improvements in computing power have yielded an astounding 200 billion factor speedup in solving Mixed Integer Optimization problems (Bertsimas et al., 2016). This is not to say that one should be contented with *any* integer programming formulation, as not all formulations are created equal. As is demonstrated in this dissertation, thinking critically about the strength of the derived formulation remains essential to expeditiously attaining a globally optimal solution.

Rather, paraphrasing a conversation I once had with Andreas Buja, we cannot allow the perceived computational constraints of the present day to overly restrict the imagination. What seems infeasible today may be feasible tomorrow, or even today if we are clever enough about it.

APPENDIX A

A.1. Summary Statistics and Percentages Missing for Covariates Used in Matching

In Table 9, we list the means and standard deviations for our non-binary covariates. Table 10 gives the percentages of ones for the binary covariates. Table 11 lists the percentages of missing values for the 13 covariates with missing data. All tables provide summaries within the ICU and hospital ward group before matching, both in our original population and in our study population defined through the solution to the maximal box problem described in Section 2.4.3.

Table 9: Means and Standard Deviations for Non-Binary Covariates Before Matching, Original Population and Study Population. The column "Tier" corresponds to the tier of importance of a given covariate as it relates to (1) the decision to admit to the hospital ward or the ICU and (2) an individual's 60 day mortality rate, as assessed by expert consultation. The next two columns are the covariate means (standard deviations) in the initial study population, and the last two columns are the covariate means (standard deviations) in the study population defined in Section 2.4.3.

		Original Population		Study Population	
Covariate	Tier	ICU	Ward	ICU	Ward
Age	1	60.1 (17.4)	55.1 (18.4)	60.56 (17.1)	55.88 (18.3)
Charlson comorbity index	1	2.52 (2.81)	2.41 (2.64)	2.43 (2.70)	2.48 (2.65)
Initial serum lactate	1	4.26 (2.98)	2.56 (1.23)	3.22 (1.24)	2.61 (0.956)
APACHE II score	1	17.7 (6.37)	13.6 (5.27)	16.9 (5.46)	13.8 (4.73)
Maximal heart rate/min	2	120 (23.6)	114 (17.8)	120 (23.0)	115 (17.8)
Maximal temperature (° F)	2	99.8 (2.97)	100.8 (2.01)	100.0 (2.90)	100.8 (2.00)
Maximal resp. rate/min	2	28.2 (9.03)	23.0 (5.77)	27.9 (8.98)	23.2 (5.90)
White blood cell count	2	14.9 (10.7)	13.0 (9.36)	14.9 (10.5)	13.0 (8.34)
Lowest systolic bp, mm Hg	2	103.2 (23.0)	107.3 (21.3)	103.4 (23.1)	108.0 (21.3)
Year of study (5-9)	2	6.73 (1.38)	6.98 (1.37)	6.65 (1.38)	7.00 (1.36)

Table 10: Percentages for Binary Covariates Before Matching, Original Population and Study Population. The column "Tier" corresponds to the tier of importance of a given covariate as it relates to (1) the decision to admit to the hospital ward or the ICU and (2) an individual's 60 day mortality rate, as assessed by expert consultation. "Exact" means that we matched exactly on that covariate. The next two columns are the percentages of ones for the covariates in the initial study population, and the last two columns are percentages of ones for the covariates in the study population defined in Section 2.4.3. Abbreviations: DNR = Do Not Resuscitate; CAD = Coronary artery disease; CHF = Congestive heart failure; COPD = Chronic obstructive pulmonary disease; ESRD = End stage renal disease.

		Original Population		Study Population	
Covariate	Tier	ICU	Ward	ICU	Ward
Female	2	56%	53%	55%	54%
Oncology	2	28%	35%	28%	36%
Transplant	2	10%	10%	12%	10%
Acute kidney infection	2	27%	15%	25%	15%
DNR order	2	4.3%	3.3%	4.0%	3.0%
Hypotension	2	32%	23%	32%	22%
Gastrointestinal infection	2	12%	12%	11%	12%
Urinary infection	2	20%	26%	21%	26%
Cellulitis	2	8.3%	13%	9.0%	13%
Bacteremia	2	22%	35%	25%	31%
Respiratory infection	2	61%	68%	63%	68%
CAD	3	11%	10%	11%	10%
CHF	3	12%	8.6%	13%	8.2%
COPD	3	8.3%	5.9%	8.4%	5.8%
Chronic liver disease	3	5.2%	3.2%	3.9%	3.1%
Chronic renal disease	3	15%	13%	15%	13%
Diabetes	3	21%	20%	22%	20%
ESRD	3	7.9%	7.9%	8.5%	8.0%
HIV	3	4.7%	3.2%	3.9%	3.6%
Hypertension	3	48%	42%	50%	42%
Cryptic septic shock	Exact	44%	10%	31%	9.0%

Table 11: Percentages of Missing Values, Original Population and Study Population. The column "Tier" corresponds to the tier of importance of a given covariate as it relates to (1) the decision to admit to the hospital ward or the ICU and (2) an individual's 60 day mortality rate, as assessed by expert consultation. The next two columns are the percentage missing in the initial study population, and the last two columns are the percentage missing in the study population defined in Section 2.4.3. Any covariate not listed here did not have missing values. Abbreviations: CAD = Coronary artery disease; CHF = Congestive heart failure; ESRD = End stage renal disease.

		Original Population		Study Population	
Covariate	Tier	ICU	Ward	ICU	Ward
Maximal heart rate/min	2	0.3%	0%	0.2%	0
Maximal temperature (° F)	2	1.0%	0.7%	0.6%	0.9%
Maximal resp. rate/min	2	0.4%	0.2%	0.2%	0.3%
White blood cell count	2	0%	0.1%	0%	0.1%
Lowest systolic bp, mm Hg	2	0.4%	0%	0.2%	0%
Bacteremia	2	51%	58%	49%	59%
Respiratory infection	2	37%	47%	34%	47%
CAD	3	1.3%	0.8%	1.4%	0.7%
CHF	3	1.2%	0.7%	0.8%	0.7%
ESRD	3	0.1%	0%	0.2%	0%

A.2. Interpolation Overlap in High Dimensions

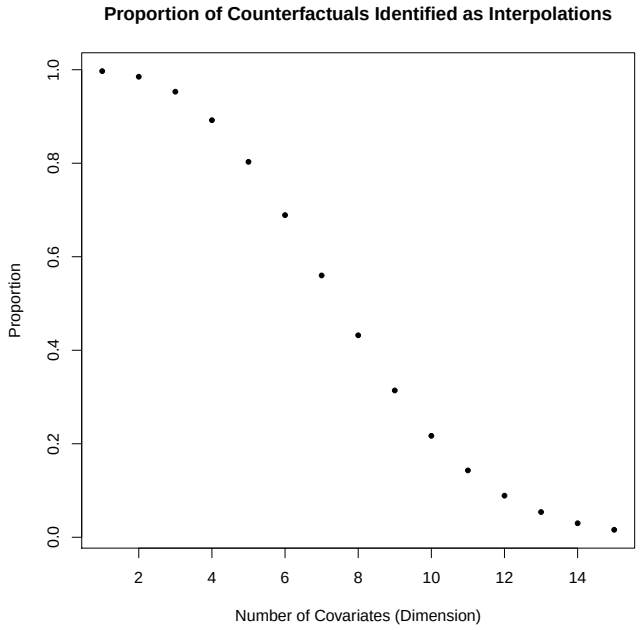

Figure 5: Proportion of individuals identified by the method of King and Zeng (2006) as lying within the area of common support as a function of covariate dimension.

To assess how attainable the interpolation overlap criterion of King and Zeng (2006) is in moderate and high dimensions we conducted a simulation study wherein treated and control individuals were truly drawn from the same multivariate distribution of increasing dimension. There were 15 simulation settings, corresponding to a covariate dimension of 1 up to 15. In the p^{th} simulation setting and for each of 1000 iterations within the p^{th} setting, we drew 1500 individuals from a p dimensional multivariate normal with a common mean and covariance matrix. After randomly assigning 750 individuals each to the treatment group and the control group, we recorded the proportion of individuals that the method of King and Zeng (2006) designated as lying within the area of interpolation overlap. For each p, we calculated the average of these proportions across iterations. The results are shown in Figure 5. As the figure displays, large percentages are identified as having counterfactuals which are estimable through interpolation in low dimensions, an occurrence one would expect since the covariates of the treated and control groups have the same joint distribution. Unfortunately,

one sees a marked decrease in the percentage of individuals lying in the area of interpolation overlap for even a moderate number of covariates despite the fact that the treated and control units are drawn from the same joint distribution. By $p = 15$, under 2% of individuals are identified as lying in the area of common support.

A.3. An Extension of the Maximal Box Problem

We now present a generalization of the maximal box problem proposed in Eckstein et al. (2002) and discussed in Section 2.4. Suppose one has a finite collection of vectors $\{\mathbf{x}_j\}, j = 1, ..., N$, that can be partitioned into two disjoint sets of "positive" points, $\mathcal{X}^+$ and "negative" points, $\mathcal{X}^-$. The generalized maximal box problem aims to find the lower and upper boundaries of a box, $[\tilde{\boldsymbol{\ell}}, \tilde{\mathbf{u}}]$, such that the corresponding box contains the maximal number of points in $\mathcal{X}^+$ while containing a fixed number C of the points in $\mathcal{X}^-$. Explicitly, $[\tilde{\boldsymbol{\ell}}, \tilde{\mathbf{u}}]$ is the arg max of the following optimization problem (GMB, for generalized maximal box):

$$\begin{aligned} \text{maximize} \quad & |[\boldsymbol{\ell}, \mathbf{u}] \cap \mathcal{X}^+| \\ \text{subject to} \quad & |[\boldsymbol{\ell}, \mathbf{u}] \cap \mathcal{X}^-| = C, \end{aligned} \tag{GMB}$$

where the notation $|A|$ denotes how many points are in set A, and $C \in \{0, 1, ..., |\mathcal{X}^-|\}$

Additional discussion of the value C is warranted. C controls how many times the maximal box is allowed to include a point which was designated to lie outside of the area of covariate overlap based on the exclusion function $\mathbf{D}(\mathbf{x}_j, \mathbf{X}, \mathbf{Z})$. Larger values of C will allow for maximal boxes that include a larger number of individuals deemed as being viable, but at the risk of including individuals for whom inference corresponds to an extrapolation of the form described in King and Zeng (2006). If one believes with absolute certainty that the elements of $\mathcal{X}^+$ are the only individuals within the area of viable covariate overlap, then allowing for $C > 0$ would result in extrapolation and C should be set to 0. Perhaps more pragmatically, if the rule used for designating a positive point is merely a means towards an end (namely, a means towards arriving at a study population wherein balance can be

attained on all covariates while overlap is present for important covariates), it is possible that allowing a small but nonzero value for C may be sensible, particularly if the maximal box returned with $C = 0$ only contains a small number of individuals. The binary nature of $\mathbf{D}(\mathbf{x}_j, \mathbf{X}, \mathbf{Z})$ is such that individuals who are designated as falling outside the area of common support may, in fact, be close to many viable individuals based on those most important covariates used to construct the maximal box. By allowing $C > 0$, one might still be able to arrive at a matching with good balance while increasing the sample size in the study population used for further analysis. Using a non-zero value of C could be thought of as recognizing that the designation of whether or not a point is inside the area of viable support will be based on guidelines which are sensible and theoretically motivated yet are not incontrovertible; see Hill et al. (2011) for further discussion of this.

A.4. Defining a Study Population through the Maximal Box Problem

In Section 2.4.3, the process of arriving at the study population used for further analysis is described in detail. Therein, we note that we designate points for exclusion from the resulting study population based on a propensity score fit on our four tier 1 covariates. As we demonstrate in Figure 1, this study population resulted in covariate overlap with respect to our most important covariates. Furthermore, this strategy resulted in a study population wherein balance could be attained on *all* of the covariates upon which we matched.

An alternative strategy would be to use the propensity score model fit on all covariates to designate whether or not a point should be excluded from the analysis while still defining the maximal box in terms of the most important covariates. When this was done with our data set, only 149 individuals (80 treated individuals, 69 control individuals) were included in the maximal box defined by our four tier 1 covariates when using the Crump et al. (2009) exclusion criterion. When using the less restrictive criterion employed within Dehejia and Wahba (1999), this number increased to 669 (347 treated, 322 control), yet over half of the individuals in our original population were nonetheless discarded. As our study population defined through the propensity score model fit on the tier 1 covariates ultimately allowed

for suitable balance to be attained, excluding such large numbers of treated and control individuals appears overly wasteful.

To see whether the exclusion of such large numbers of individuals when using the full propensity score model was a limitation of our method versus a more general issue with attaining common support in high dimensions, we also used the procedure of King and Zeng (2006) with all of our covariates. As a reminder, the method of King and Zeng (2006) does not utilize propensity scores and instead defines the area of common support as the intersection of the convex hulls of covariates for treated and control individuals. Using this definition with all of our covariates, we found that no individuals were identified as lying within the area of viable common support: no treated individuals were within the convex hull of the control units, and no control individuals were within the convex hull of the treated units.

Given these developments, we see our procedure as pursuing a goal which may be more attainable in practice: attaining overlap and balance with respect to the important covariates while seeking balance on all covariates. As such, we fit our propensity score model with respect to only the tier 1 covariates. An extreme estimated propensity score based on these important covariates then indicates that an observation is "extreme" with respect to the distributions of the most important covariates for either the treated or control groups. As mentioned in the manuscript, using this limited model to create the maximal box resulted in 1208 individuals in our subpopulation (507 ICU, 701 hospital ward). We also used the method of King and Zeng (2006) with respect to these most important covariates, and found that 1227 individuals were designated as lying within the area of common support (505 ICU, 722 hospital ward). The advantage of our method over that of King and Zeng (2006) is that the study population used for further analysis based on the maximal box is easily described and interpreted in terms of ranges of covariate values.

While we were able to attain balance in our study population, this need not be the case with other data sets. Although overlap should be attained with respect to the most important covariates imbalances may persist even after matching, particularly for covariates not used to

define the study population. One approach for dealing with this would be an iterative process wherein covariates which cannot be suitably balanced even after matching are used to define a new maximal box. One would then create a new match in this new study population, and reassess the resulting balance after matching. Iteratively refining the study population would not bias the resulting analysis; rather, the individuals to whom the inference applies would change with each iteration.

A.5. Strength of the Normal Approximation Under the Worst-Case Allocation of Potential Outcomes

Our procedure for testing the composite null hypothesis $\delta = \delta_0$ hinges upon the appropriateness of the normal distribution in approximating the randomization distribution of the estimated average treatment effect, $\hat{\delta}$. In Section 2.5.1 we prove asymptotic normality of $\hat{\delta}$ under mild conditions, but we would like to see how well the normal approximation holds for the example at hand. Our optimization problem returns vectors of potential outcomes under treatment and control corresponding to the worst-case variance of $\hat{\delta}$. To assess whether the normal approximation is reasonable for this allocation, we perform a Monte Carlo simulation to generate samples from the *true* randomization distribution of $\hat{\delta}$ using the worst-case potential outcomes for inference within the full match described in Section 2.4.3. Randomization occurs independently across strata. In stratum i, we randomly generate a vector $\mathbf{Z}_i$ with m_i ones and $n_i - m_i$ zeroes, corresponding to the assignment to treatment and control respectively. The observed outcome for individual j in stratum i is then $R_{ij} = r_{Tij}Z_{ij} + r_{Cij}(1 - Z_{ij})$. This yields the estimated average treatment effect in stratum i: $\hat{\delta}_i = \sum_{j=1}^{n_i} (Z_{ij}R_{ij}/m_i - (1 - Z_{ij})R_{ij}/(n_i - m_i))$. Finally, we form $\hat{\delta} = \sum_{i=1}^{I}(n_i/N)\hat{\delta}_i$.

Figure 6 shows the resulting randomization distribution under this worst-case allocation of potential outcomes. We first note that the distribution is centered at $\mathbb{E}[\hat{\delta}] = 0$, as we are testing the composite null that $\delta = 0$. Furthermore, we note that both the histogram and the normal quantile plot indicate that the randomization distribution can be well approximated by a normal distribution.

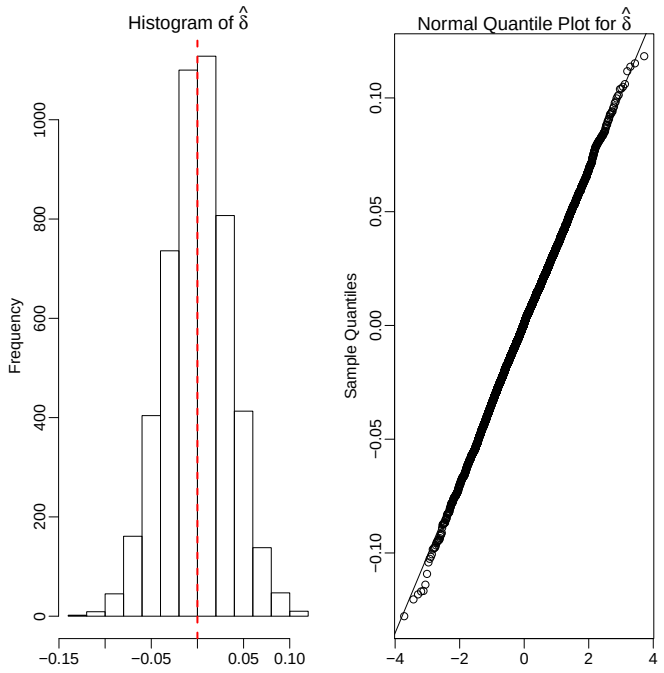

Figure 6: Randomization distribution of $\hat{\delta}_i$ under the worst-case allocation of potential outcomes returned by the optimization procedure described in Section 2.5.2. The dotted vertical line in the histogram corresponds to average treatment effect under the null, 0.

A.6. A Comparison of Standard Errors

We compare the value of the standard error used in conducting our hypothesis test, 3.67%, to standard errors associated with three other hypothesis tests:

1. A simple two-sample test for differences in proportion assuming *iid* draws from the two populations of interest. That is, we do not take the stratification into account, and find the standard errors under a biased analysis assuming two *iid* samples. $\mathbf{SE}(\hat{\delta}) = \mathbf{2.26}\%$

2. A test of $\delta = 0$ using the Mantel-Haenszel risk difference estimator (Greenland and Robins, 1985). We assume independent sampling across strata, that we have independent draws within each stratum from each of the two groups being compared, and that there is a common treatment effect across all strata ($\delta_i = \delta_0 \; \forall i$). Given the nature of our stratification (a large number of strata with a limited number of observations in each stratum), we use the variance estimator of Sato et al. (1989) which is consistent

under sparse stratification. $\mathbf{SE}(\hat{\delta}) = \mathbf{2.59}\%$

3. A test of Fisher's sharp null under our stratification ($r_{Tij} = r_{Cij}\ \forall i, j$).
   $\mathbf{SE}(\hat{\delta}) = \mathbf{2.67}\%$

Note that the first procedure has a larger effective sample size than that of our idealized stratified experiment since it assumes *iid* draws from two populations, while the other two estimators account for the stratification at hand. Given our estimated ATE of 4.3%, we would still fail to reject the null with any of these alternate standard errors. None of these three alternate procedures have size α for all alignments of potential outcomes within the composite null hypothesis, as our procedure finds the actual worst-case standard error over all elements of the composite null.

Perhaps most interesting is the comparison of our maximal standard error to that attained under Fisher's sharp null. Clearly, the sharp null of no treatment effect is an element of the composite null $\delta = 0$. To see why the variances are so different, note that under the sharp null, any stratum with $R_{ij} = R_{ik}$ for all $j, k \in \{1, ..., n_i\}$ yields $\text{var}(\hat{\delta}_i) = 0$, since the missing potential outcome for each individual must equal the observed value for that individual. Under the general composite null $\delta = 0$, we can arrange the potential outcomes across strata in such a way that the variances of stratum-specific ATEs are positive even if $R_{ij} = R_{ik}$ for all $j, k \in \{1, ..., n_i\}$. As a simple illustration of this, consider testing the null of $\delta = 0$ under both the sharp null and under the composite null with two strata. In stratum 1, suppose $R_{11} = R_{12} = 1$, while in stratum 2 suppose $R_{21} = R_{22} = 0$, where without loss of generality the first individual in each matched set received the treatment. If we assume the sharp null holds, $r_{C11} = 1$, $r_{T12} = 1$, $r_{C21} = 0$ and $r_{T22} = 0$. Within each of these strata, the variance of the stratum-specific average treatment effect is 0. On the other hand, we can also satisfy the composite null $\delta = 0$ by setting $r_{C11} = 1$, $r_{T12} = 0$, $r_{C21} = 0$, $r_{T22} = 1$, which would yield $\text{var}(\hat{\delta}_i) > 0$ for $i = 1, 2$. Neyman's null offers more flexibility for the optimization problem, which is why we see the discrepancy between the standard errors from our procedure with those under the sharp null. Strata where $R_{ij} = R_{ik}$ for all

$j, k \in \{1, ..., n_i\}$ occur regularly in our example (182 out of 312 strata), which explains the magnitude of the difference between the two standard errors.

A.7. Formulating the Maximal Variance Problem

We now introduce notation for maximizing the variance within a composite null through integer programming. While many different formulations are possible, the one we choose explicitly avoids symmetry by having each decision variable correspond to a unique distribution on the contribution to the overall estimated average treatment effect from a given stratum. As discussed in Margot (2010), the avoidance of symmetry is crucial in formulating an integer program that can be solved in a reasonable amount of computation time.

Let $\mathcal{T}_i^{zr} = \{j : Z_{ij} = z, R_{ij} = r\}$, $(z, r) \in \{0, 1\}^2$, $i \in \{1, ..., I\}$, denote the four possible partitions of indices of individuals in stratum i into sets based on the values of their treatment assignment and observed response. Within each set, all members share the same value for either r_{Tij} or r_{Cij}. For example, if $j, k \in \mathcal{T}_i^{01}$, then $r_{Cij} = r_{Cik} = 1$, yet the values of r_{Tij} and r_{Tik} are unknown. $|\mathcal{T}_i^{zr}|$ can be thought of as the value in cell (z, r) of a 2×2 table that counts the number of individuals for each combination of (z, r) in stratum i. To tie notation together, we have that $|\mathcal{T}_i^{11}| + |\mathcal{T}_i^{10}| = m_i$, and $|\mathcal{T}_i^{01}| + |\mathcal{T}_i^{00}| = n_i - m_i$

There are 2^{n_i} possible sets of potential outcomes in stratum i that are consistent with the observed data. As we now show, we need not consider all 2^{n_i} possible combinations of potential outcomes, but rather only those which correspond to unique distributions of $\hat{\delta}_i$. This is beneficial as the 2^{n_i} possible sets of potential outcomes in stratum i only yield $\prod_{(z,r)\in\{0,1\}^2}(|\mathcal{T}_i^{zr}| + 1)$ unique distributions for $\hat{\delta}_i$. This is demonstrated in Rigdon and Hudgens (2014, Section 3), and the argument is reproduced here. Consider $\mathcal{T}_i^{00}$. The potential outcomes under treatment are unknown in this set; however, since the potential outcomes under control are the same for all individuals, the possible allocations of $\{r_{Tij} : j \in \mathcal{T}_i^{00}\}$ only result in $|\mathcal{T}_i^{00}| + 1$ non-exchangeable distributions. These are attained by setting $\{r_{Tij} : j \in \mathcal{T}_i^{00}\}$ equal to any one of the ordered vectors

$(0,0,...,0),(1,0,...,0)$,..., and $(1,1,...,1)$. The same argument holds for the other three sets of indices in stratum i, thus completing the proof. Note further that since n_i and m_i are fixed across randomizations and $\min\{m_i, n_i - m_i\} = 1$, we have that $\prod_{(z,r)\in\{0,1\}^2}(|\mathcal{T}_i^{zr}| + 1) = 2 \times \max\left\{\prod_{(r)\in\{0,1\}}(|\mathcal{T}_i^{0r}| + 1), \prod_{(r)\in\{0,1\}}(|\mathcal{T}_i^{1r}| + 1)\right\}$.

It would seem as though we must consider $\prod_{i=1}^{I}\prod_{(z,r)\in\{0,1\}^2}(|\mathcal{T}_i^{zr}|+1)$ different distributions for the estimated average treatment effect in our optimization problem. Fortunately, two facts allow us to consider a much smaller number of variables. First, there is independence between strata which allows us to sum stratum-wise variance contributions together to arrive at the overall variance of the estimated average treatment effect. Second, many of the stratum-specific 2×2 tables are observed multiple times across strata. As an example, our full match returned 312 strata, of which there were only 48 unique tables.

In light of these facts, we introduce notation to facilitate the solution of our optimization problem. Let $\mathcal{C}_i = (|\mathcal{T}_i^{00}|, |\mathcal{T}_i^{01}|, |\mathcal{T}_i^{10}|, |\mathcal{T}_i^{11}|)$ be the observed counts of the 2×2 table for stratum i. $\mathfrak{C} = \{\mathcal{C}_1, ..., \mathcal{C}_I\}$ is a (multi)set, where the number of unique elements equals the number of unique 2×2 tables observed in the data and is typically much smaller than I. Let S be the number of unique tables and let $s \in \{1, ..., S\}$ index the unique tables. Define $\mathcal{I}(i)$ to be a function returning the index of the unique table corresponding to the table observed in stratum i. Hence, $\mathcal{I}(i) = \mathcal{I}(\ell)$ if and only if $\mathcal{C}_i = \mathcal{C}_\ell$. Let $M_s = |\mathcal{I}^{-1}(s)|$ be the number of strata where unique table s was observed, and let $\tilde{n}_s = n_b$ and $\tilde{m}_s = m_b$ be the number of total units and treated units respectively in unique table s for any $b \in \mathcal{I}^{-1}(s)$. Let P_s be the number of allowed non-exchangeable potential outcomes for unique table s, and let $\{[\mathbf{r}_{T[sp]}, \mathbf{r}_{C[sp]}]\}, p \in \{1, ..., P_s\}$ be the set of allowed potential outcome allocations for unique table s. Finally, let $\delta_{[sp]j} = r_{T[sp]j} - r_{C[sp]j}$, and let $\Delta_{[sp]} = \sum_{j=1}^{\tilde{n}_s}\delta_{[sp]j}$.

Define $\nu_{[sp]}$ as:

$$\nu_{[sp]} = \frac{\tilde{n}_s^2}{N^2}\left(\frac{S^2_{T[sp]}}{\tilde{m}_s} + \frac{S^2_{C[sp]}}{\tilde{n}_s - \tilde{m}_s} - \frac{S^2_{\delta[sp]}}{\tilde{n}_s}\right),$$

where $S^2_{T[sp]} = \sum_{j=1}^{\tilde{n}_s}(r_{T[sp]j} - \bar{r}_{T[sp]})^2/(\tilde{n}_s - 1)$, $S^2_{C[sp]} = \sum_{j=1}^{\tilde{n}_s}(r_{C[sp]j} - \bar{r}_{C[sp]})^2/(\tilde{n}_s - 1)$, and $S^2_{\delta[sp]} = \sum_{j=1}^{\tilde{n}_s}(\delta_{[sp]j} - \bar{\delta}_{[sp]})^2/(\tilde{n}_s - 1)$. $\nu_{[sp]}$ then represents the variance of the contribution to the overall estimated average treatment effect from table s under potential outcome allocation p. Let $\boldsymbol{\nu} = [\nu_{[11]}, ..., \nu_{[sp]}]$.

Let $x_{[sp]}$ be an integer decision variable denoting how many times the set of potential outcomes p that is consistent with unique table s is observed in the data, $s \in \{1, ..., S\}$, $p \in \{1, ..., P_s\}$, and let $\mathbf{x} = [x_{[11]}, .., x_{[SP_s]}]$. $\nu_{[sp]}x_{[sp]}$ represents the contribution to the overall variance of the test statistic if the p^{th} set of potential outcomes in unique table s is observed $x_{[sp]}$ times, and $\boldsymbol{\nu}^T\mathbf{x}$ represents the overall variance across all unique tables and potential outcomes that are observed in the data. $\sum_{p=1}^{P_s} x_{[sp]}$ is how many times the s^{th} unique table was observed in the data, which through our definition of M_s results in the constraint that $\sum_{p=1}^{P_s} x_{[sp]} = M_s \ \forall s$. Finally, we force the resulting optimal solution to have an allocation of potential outcomes such that the null hypothesis in question is satisfied (that is, $\boldsymbol{\delta} \in \mathcal{D}_{\delta_0}$) through an additional constraint. Given our definition of $\Delta_{[sp]}$, the constraint that the null must be true can be written as $\sum_{s=1}^{S}\sum_{p=1}^{P_s} \Delta_{[sp]}x_{[sp]} = N\delta_0$. Finding the maximal variance over all $\boldsymbol{\delta} \in \mathcal{D}_{\delta_0}$ can then be written as the following linear integer program (MV, for maximal variance):

$$
\begin{aligned}
\underset{\mathbf{x}}{\text{maximize}} \ & \boldsymbol{\nu}^T\mathbf{x} \\
\text{subject to} \ & \sum_{p=1}^{P_s} x_{[sp]} = M_s \ \ \forall s \\
& \sum_{s=1}^{S}\sum_{p=1}^{P_s} \Delta_{[sp]}x_{[sp]} = N\delta_0 \\
& x_{[sp]} \in \mathbb{Z} \ \ \forall s, p \\
& x_{[sp]} \geq 0 \ \ \forall s, p,
\end{aligned}
\tag{MV}
$$

where $\mathbb{Z}$ denotes the set of integers.

For a given δ_0, we can then use the objective value of (MV) at the optimal value $\mathbf{x} = \mathbf{x}^*_{\delta_0}$ to perform a hypothesis test of $\delta = \delta_0$. This procedure is *not* conservative for testing the composite null, as the maximal variance is attained by a member of the composite null: there is an allocation of potential outcomes that satisfies the null, aligns with the observed data, and has this variance. Confidence intervals can then be attained by inverting tests on a sequence of values of $\{\delta_0\}$. To aid in finding the endpoints of this interval, we can start with a Wald-type confidence interval found by finding the maximal variance at $\delta_0 = \hat{\delta}$, forming an interval of the form $\hat{\delta} \pm z_{1-\alpha/2}\sqrt{\nu^T \mathbf{x}^*_{\hat{\delta}}}$, and then refining the endpoints through a series of tests for values of δ_0 near the endpoints of the Wald interval.

APPENDIX B

B.1. Balance on Observed Covariates in Our Motivating Example

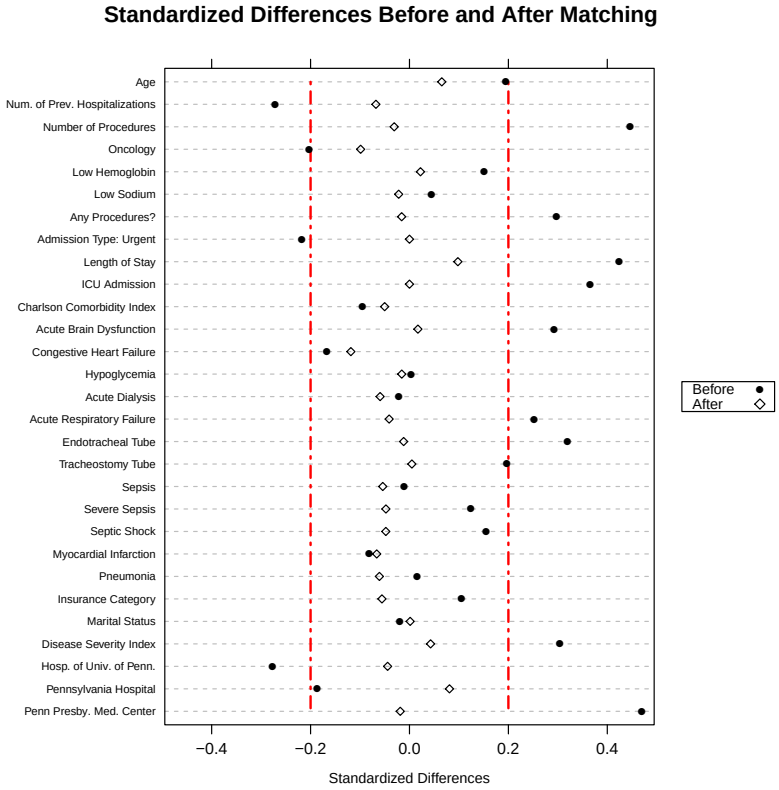

Figure 7: Covariate Imbalances Before and After Matching. The dotplot (a Love plot) shows the absolute standardized differences without matching, and after conducting a matching with a variable number of controls. The vertical dotted line corresponds to a standardized difference threshold of 0.2, which is often regarded as the maximal allowable absolute standardized difference (Rosenbaum, 2010). As one can see, marked imbalances existed between the two populations before matching. All standardized differences were below 0.2 after matching, and most covariates saw substantial improvements in balance through matching.

B.2. Usage of Risk Differences and Risk Ratios

The risk difference and risk ratio are two measures of the causal effect of an intervention on a binary outcome. A common viewpoint taken in the statistics literature is that the appropriateness of using the risk ratio (also called the relative risk) versus the risk difference depends on the scale of the problem, with certain measures being appropriate for certain inferences. This is discussed in Hernán and Robins (2016) in the following paragraph:

> Each effect measure may be used for different purposes. For example, imagine a large population in which 3 in a million individuals would develop the outcome if treated, and 1 in a million individuals would develop the outcome if untreated. The causal risk ratio is 3, and the causal risk difference is 0.000002. The causal risk ratio (multiplicative scale) is used to compute how many times treatment, relative to no treatment, increases the disease risk. The causal risk difference (additive scale) is used to compute the absolute number of cases of the disease attributable to the treatment. The use of either the multiplicative or additive scale will depend on the goal of the inference. (Hernán and Robins, 2016, pages 7-8)

Of course, the converse can be true: if 85% develop the outcome if treated and 80% develop the outcome if not treated, the risk ratio is then 1.0625 while the risk difference is 0.05. Grieve (2003) provides additional discussion of these two estimands, noting that in deciding which estimand to use one must consider "whether interest is centered on absolute or relative effects, and the extent to which those who are to use them understand them" (Grieve, 2003, page 88).

The summary measure chosen can also affect the extent to which a study's findings influence future action. Misselbrook and Armstrong (2001) note that when deciding whether or not to take a proposed treatment the percentage of individuals who end up agreeing to take the treatment can vary substantially depending on whether the benefits of a treatment are presented in the form of a risk ratio or a risk difference. Forrow et al. (1992) note that the manner in which information on a causal effect is presented can affect not only how likely patients are to take a recommended treatment, but also how likely a doctor is to prescribe a treatment in the first place.

Poole (2010) states that in epidemiology, it has been treated as a seemingly self-evident truth that "relative effect measures should be used to assess causality and that absolute measures should be used to assess impact." (Poole, 2010, page 3). An early defense of this

stance can be found in the work of Cornfield et al. (1959) on smoking and lung cancer:

> Both the absolute and the relative measures serve a purpose. The relative measure is helpful in (1) appraising the possible noncausal nature of an agent having an apparent effect; (2) appraising the importance of an agent with respect to other possible agents inducing the same effect; and (3) properly reflecting the effects of disease misclassification or further refinement of classification. The absolute measure would be important in appraising the public health significance of an effect known to be causal. (Cornfield et al., 1959)

Both Poole (2010) and Ding and Vanderweele (2014) refute the superiority of the risk ratio to the risk difference in making causal claims, presenting examples where the use of evidence presented by the risk difference exhibits much stronger robustness to unmeasured confounding than evidence presented by the risk ratio, thus aiding in discovering causal effects.

In the clinical trials literature, both effect measures are viewed as having their own relative merits and downsides. Schechtman (2002) takes a pragmatic approach and suggests that in order to paint a clearer picture of the treatment effect, one should report both the estimated risk difference and risk ratio. See Cook and Sackett (1995), Jaeschke et al. (1995), and Sinclair and Bracken (1994) for further discussion of this matter.

B.3. Simulation Studies for Computation Time

Our methodology can, for the purposes of computation time, be thought of as containing three components with worst case complexities as follows:

1. Defining groups of symmetric tables: $O(I^2)$

2. Defining constants and constraints for unique tables:
   $O\left(S + \sum_{s=1}^{S}(\tilde{n}_s - 1)\prod_{(z,r,d)\in\{0,1\}^3}(|\mathcal{T}_s^{zrd}| + 1)^2\right)$

3. Solution of integer program: $\mathcal{NP}$-hard

For the first component, the total number of matched sets plays a role in determining computation time as in formulating the problem, we must sort the individual matched sets into symmetry groups corresponding to uniquely observed tables. The second component is affected not only by the number of uniquely observed tables, but also the number of observations in a table and the cells of said table. As discussed in Section 3.4, each table s yields at most $\prod_{(z,r,d)\in\{0,1\}^3}(|\mathcal{T}_s^{zrd}|+1)^2$ unique distributions, while for a sensitivity analysis there are $\tilde{n}_s - 1$ alignments of the unmeasured confounders to be considered for each distribution. These unique contributions correspond to variables in our optimization problem. The number of variables is also influenced by assumptions made on the potential outcomes, as assumptions eliminate the need to consider certain possible values for the unobserved potential outcomes.

The simulation studies presented herein provide further insight into various aspects of problem (P1) which can affect the solution of the integer program itself (component 3), as this is the only $\mathcal{NP}$-hard endeavor and hence may, in theory, lead to unpredictable computation time. Unless otherwise stated, all of the simulations presented are modifications of the same basic set up. In each of 1000 iterations we sample I matched sets from the strata in our motivating example from Section 3.1.2. Each iteration has strata ranging in size from 2 to 21, and each data set has an average of roughly $8 \times I$ individuals within it. Large strata affect computation time, as they result in larger numbers of non-exchangeable potential outcome allocations within a stratum and fewer duplicated 2×2 tables in the data. In our data set, 25% of the strata had one acute rehabilitation individual and 20 home with home health services patients. Treated and control individuals are assigned an outcome of "1" with probability p_T and p_C respectively.

In each iteration, we test a null on the causal risk difference, $\delta = \delta_0$. We test the stated null with a two-sided alternative at level of unmeasured confounding Γ. We record the required time for the optimization problem itself for each simulation. Simulations were conducted

on a desktop computer with a 3.40 GHz processor and 16.0 GB RAM. The `R` programming language was used to formulate the optimization problem, and the `R` interface to the `Gurobi` optimization suite was used to solve the optimization problem.

B.3.1. Increasing the Number of Matched Sets

In this simulation, we fix $p_T = 0.75, p_C = 0.25, \Gamma = 2, \delta_0 = 0.2$, and conduct 1000 iterations at $I = 7, 13, 65, 125, 625$. As Figure 8 demonstrates, the time for the optimization routine itself appears to increase with I, the number of matched sets. Figure 8 also demonstrates that time is increasing with the average number of variables in the corresponding optimization problem.

To demonstrate that the role that I plays is only indirect (through its effect on the number of variables in the optimization problem), we also present a simulation study with matched sets of size three. We will focus on the effect ratio in this simulation study. Each set consists of three individuals, one encouraged to take the treatment and the other two encouraged to take the control. For each individual, the probability of compliance with the assigned treatment is set to 0.9. We set $p_T = 0.75$ and $p_C = 0.25$ based on which treatment the individual actually received. We set $\Gamma = 2$ and $\lambda_0 = 0.2$, and conduct 1000 iterations with $I = 25, 50, 250, 500, 2500, 5000, 25000, 50000, 250000$. In the corresponding inference, we do not assume that the exclusion restriction holds. We also do not assume monotonicity holds, nor do we assume a known direction of effect.

Figure 9 shows that as I increases the time required for only solving the optimization problem initially increases, but then begins to level off. The reason for this is also demonstrated in the figure: as I increases, the average number of variables in the optimization problem appears to be approaching an asymptote, rather than continually increasing. This is because under the assumptions used for the performed inference, the maximal number of unique allocations of unobserved potential outcomes and unmeasured confounders that must be considered is 4384, calculated using the formula $\sum_{s=1}^{S}(\tilde{n}_s - 1)\prod_{(z,r,d)\in\{0,1\}^3}(|\mathcal{T}_s^{zrd}| + 1)^2 =$

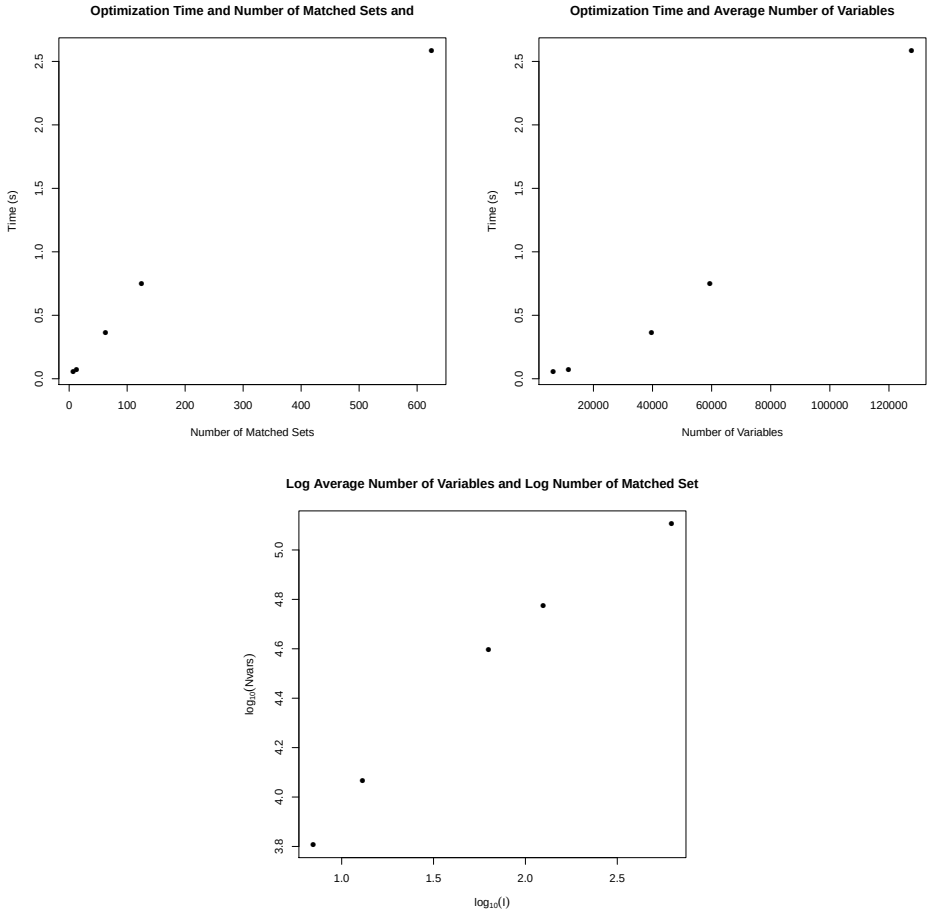

Figure 8: (Top-left) Optimization time and the number of matched sets; (top-right) optimization time and the number of optimization variables; and (bottom) log number of matched sets and log number of optimization variables.

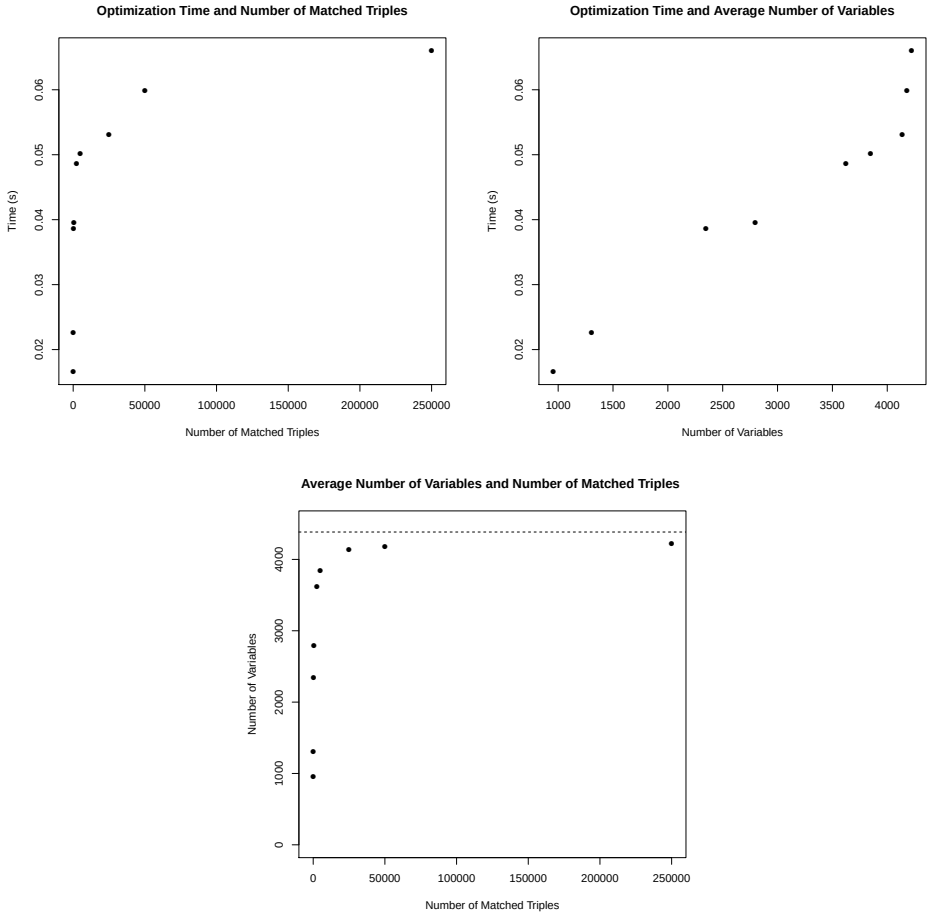

Figure 9: (Top-left) Optimization time and the number of matched triples; (top-right) optimization time and the average number of variables; and (bottom) average number of variables and the number of matched triples.

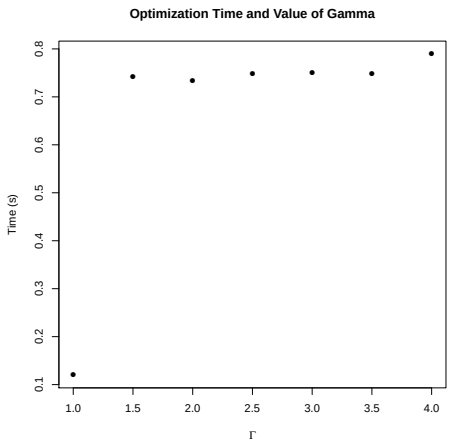

Figure 10: Optimization time and value of Γ.

$2 \times (4 \times 3^2 \times 2^2 + 32 \times 2^6)$. This illustrates one of the key advantages of our formulation: by expressing the problem in terms of unique contributions to the test statistic we greatly enhance the scalability of our method, particularly when the matched sets are of limited size. In fact, the average computation time for the optimization problem was under a tenth of a second for all values of I in this simulation setting.

B.3.2. Increasing the Value of Γ

In this simulation we fix $p_T = 0.75, p_C = 0.25, I = 125, \delta_0 = 0.2$, and conduct 1000 iterations at each of $\Gamma = 1, 1.5, , ..., 3.5, 4$. We see in Figure 10 that while there is a substantial increase in solution time when going from $\Gamma = 1$ to $\Gamma > 1$, the solution time is roughly constant at all values of $\Gamma > 1$ tested. $\Gamma = 1$ corresponds to an integer linear program while any $\Gamma > 1$ is an integer quadratic program, which accounts for the initial jump. Increasing Γ further does not change the fact that it is an integer quadratic program, nor does it increase the average number of variables in the optimization problem; rather, it changes the values of the constants associated with each of the variables in the objective function.

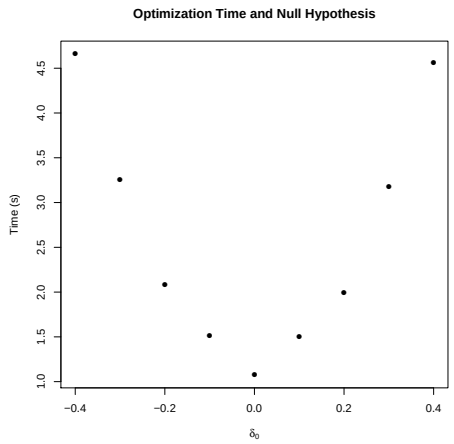

Figure 11: Optimization time and null hypothesis being tested. The true risk difference was set to zero throughout

B.3.3. Altering the Hypothesized Risk Difference

In this simulation we fix $p_T = 0.5, p_C = 0.5, I = 125, \Gamma = 2$, and conduct 1000 iterations at each of $\delta_0 = -0.4, -0.3, ..., 0.3, 0.4$. As we see in Figure 11, average solution time is shortest when the true risk difference is closest to the hypothesized risk difference, and increases as the hypothesized risk difference moves away from the truth in either direction. Note that both the number of variables and the number of constraints in the optimization problem remain constant on average as the hypothesized risk difference varies, meaning that neither can explain the difference in solution times. As δ_0 moves further away from the true risk difference the average number of *feasible solutions* decreases, as the discrepancy between the observed potential outcomes and the null hypothesis affords less and less flexibility to the allocation of the unobserved potential outcomes. This can, in turn, make the corresponding integer program more difficult to solve.

B.3.4. Jointly Altering the Outcome Prevalence Under Treatment and Control

In this simulation we fix $I = 125, \Gamma = 2, \delta_0 = 0$, and conduct 1000 iterations at each of $[p_C, p_T] = [0.05, 0.15], [0.15, 0.25], ..., [0.85, 0.95]$. Hence, the distance between the null hypothesis and the true risk difference remains constant at 0.1. In Figure 12, we see that

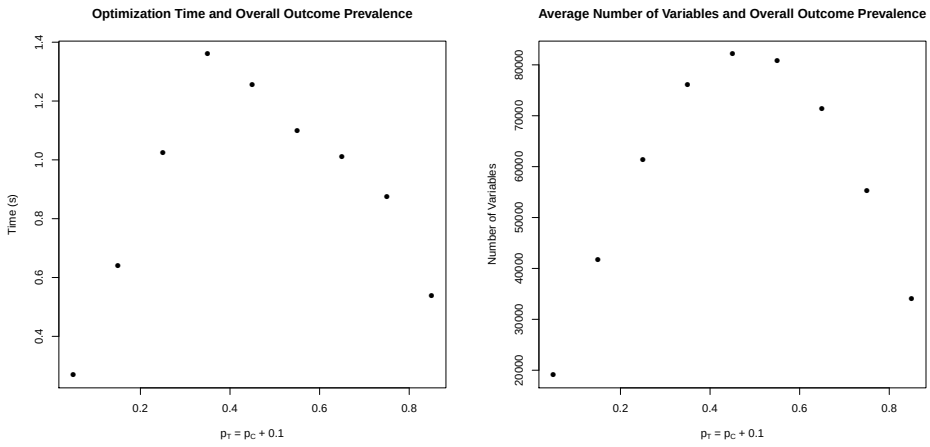

Figure 12: (Left) Optimization time and overall outcome prevalence; and (right) number of variables and outcome prevalence.

simulation time is greatest when the outcomes have the highest variance (i.e., when the treated and control prevalences are closest to 0.5), but drop off when the outcome becomes either rarer or highly prevalent. Figure 12 also shows the relationship between the number of variables and the outcome prevalence. The number of unique contributions to the overall test statistic from a given unique table (i.e. the number of variables) is maximized when the outcome prevalences are closest to 0.5, which accounts for the observed computation time pattern.

B.3.5. Separately Altering the Outcome Prevalence Under Treatment and Control

In our first simulation, we fix $p_C = 0.1, I = 125, \Gamma = 2, \delta_0 = 0$, and conduct 1000 iterations at each of $p_T = 0.1, ..., 0.9$. In Figure 13, we see that the outcome prevalence under treatment affects computation time by increasing the number of variables in the optimization problem.

Next, we fix $p_T = 0.9, I = 125, \Gamma = 2, \delta_0 = 0$, and conduct 1000 iterations at each of $p_C = 0.1, ..., 0.9$. In Figure 14, we see that the outcome prevalence under control affects computation time by increasing the average number of variables in the optimization problem. Note that altering the prevalence under control has a more drastic effect on the number of variables (and thus, on the overall computation time) than altering the prevalence under

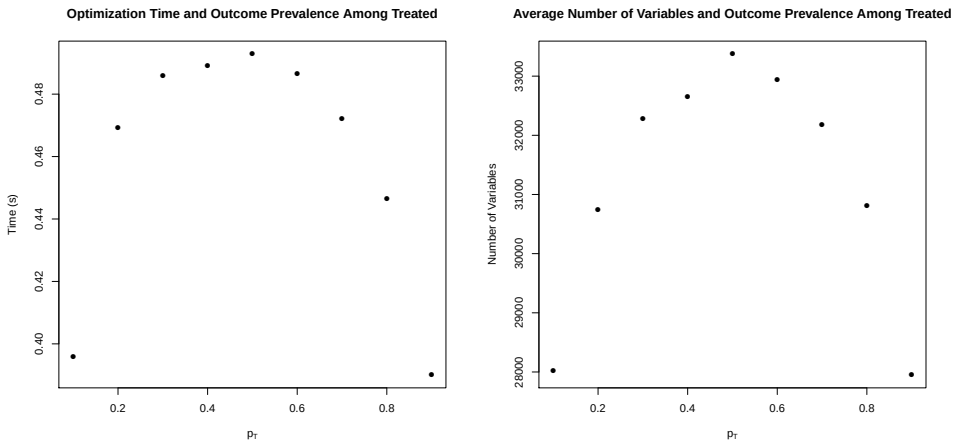

Figure 13: (Left) Optimization time and outcome prevalence under treatment; and (right) number of variables and outcome prevalence under treatment.

treatment, as the matched sets used in our simulation study each have one treated unit and one or more (up to 20) control units. In turn, heterogeneity among control units within a given matched set allows for many more possible contributions to the overall test statistic (variables), particularly in matched sets with large numbers of control units. When altering the prevalence for the treated units, since there is only one treated unit per matched set an event prevalence for treated units closer to 0.5 only increases the number of variables in the optimization problem by making it less likely that two matched sets with the same observed table for the control units also have the same observed response for their respective treated unit.

B.3.6. Assessing Avoidance of Symmetry

At $\Gamma = 1$, we compare computation time of our formulation, formulation (P1), for the causal risk difference with that of an equivalent binary programming formulation. We first present this alternate formulation. Let v_{ij} be the unobserved potential outcome for each individual. That is, $v_{ij} = r_{Cij}$ if $Z_i = 1$, and $v_{ij} = r_{Tij}$ if $Z_i = 0$. When conducting inference assuming no unmeasured confounders ($\Gamma = 1$), we aim to find the worst-case variance among the set of unobserved potential outcomes such that the null is satisfied, a problem which can be expressed as a quadratic form involving the unobserved potential outcomes and other

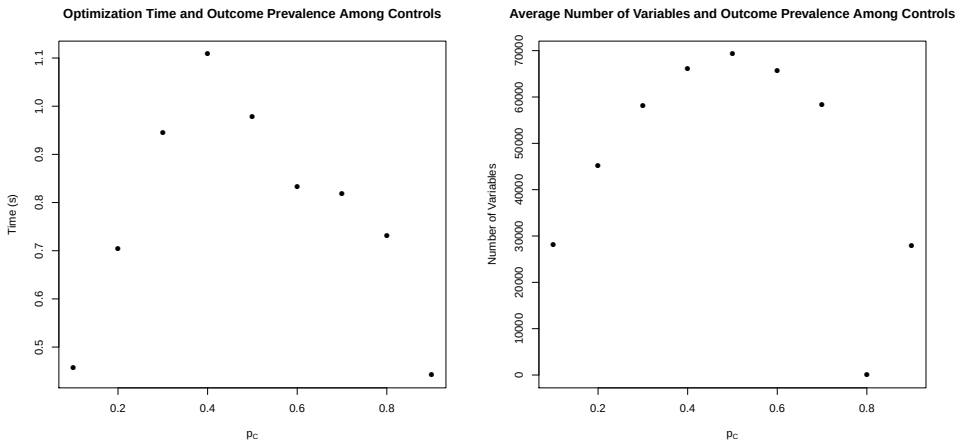

Figure 14: (Left) Optimization time and outcome prevalence under control; and (right) number of variables and outcome prevalence under control.

constants known at the time of the optimization. Using the methods of Glover and Woolsey (1974) for converting a quadratic binary program into a linear binary program, we can express the problem as:

$$\begin{aligned}
\text{maximize} \quad & \sum_{i=1}^{I}\sum_{j=1}^{n_i} p_{ij}v_{ij} + 2\sum_{i=1}^{I}\sum_{j<k\leq n_i} p_{ijk}w_{ijk} + c \qquad \text{(AP1)} \\
\text{subject to} \quad & \sum_{i=1}^{I}\sum_{j=1}^{n_i}(2Z_{ij}-1)v_{ij} = -N\delta_0 + \sum_{i=1}^{I}\sum_{j=1}^{n_i}(2Z_{ij}-1)R_{ij} \\
& v_{ij} \in \{0,1\} \;\; \forall i,j \\
& w_{ijk} \leq v_{ij}, v_{ik} \;\; \forall i,j,k \\
& v_{ij} + v_{ik} - w_{ijk} \leq 1 \;\; \forall i,j,k
\end{aligned}$$

We now define p_{ij}, p_{ijk} and c. Let $\mathbf{H}^{(i)}$ be an $n_i \times n_i$ symmetric matrix with diagonal elements $(n_i^2 - n_i)/N^2$ and off diagonal elements are $-n_i/N^2$, and define the following vectors. Let $\mathbf{A}^{(i)}$ be an $n_i \times n_i$ diagonal matrix with diagonal entries $1/(Z_{i,n_i}(2-n_i)+n_i-1)$, and let $\mathbf{B}^{(i)}$ be an $n_i \times n_i$ diagonal matrix with diagonal entries $1/((1-Z_{i,n_i})(2-n_i)+n_i-1)$. We

can then write $\text{var}(T(\delta_0))$ as a sum of stratum-specific quadratic forms:

$$\begin{aligned}\text{var}(T(\delta_0)) &= \sum_{i=1}^{I} \left([\mathbf{A}^{(i)}\mathbf{R}_i + \mathbf{B}^{(i)}\mathbf{v}_i]^T \mathbf{H}^{(i)} [\mathbf{A}^{(i)}\mathbf{R}_i + \mathbf{B}^{(i)}\mathbf{v}_i] \right) \\ &= \sum_{i=1}^{I} \left(\mathbf{v}_i^T \mathbf{B}^{(i)} \mathbf{H}^{(i)} \mathbf{B}^{(i)} \mathbf{v}_i + 2\mathbf{v}_i^T \mathbf{B}^{(i)} \mathbf{H}^{(i)} \mathbf{A}^{(i)} \mathbf{R}_i + \mathbf{R}_i^T \mathbf{A}^{(i)} \mathbf{H}^{(i)} \mathbf{A}^{(i)} \mathbf{R}_i \right)\end{aligned}$$

Let $p_{ij} = (\mathbf{B}^{(i)}\mathbf{H}^{(i)}\mathbf{A}^{(i)}\mathbf{R}_i)_j + (\mathbf{B}^{(i)}\mathbf{H}^{(i)}\mathbf{B}^{(i)})_{jj}$, $p_{ijk} = (\mathbf{B}^{(i)}\mathbf{H}^{(i)}\mathbf{B}^{(i)})_{jk}$, and $c = \sum_{i=1}^{I} \mathbf{R}_i^T \mathbf{A}^{(i)} \mathbf{H}^{(i)} \mathbf{A}^{(i)} \mathbf{R}_i$, we recover the required constants for finding the maximal variance of the causal risk difference.

Rather than having decision variables for each possible variance contribution, this formulation has binary decision variables for the missing potential outcome for each individual. A formulation of this sort yields a highly symmetric problem, as any pair of individuals in a given stratum with $[Z_{ij}, R_{ij}] = [Z_{ik}, R_{ik}]$ are exchangeable. For example, if individual j and k in stratum i both received the control and had an outcome of 0, then $r_{Tij} = 1, r_{Tik} = 0, u_{ij} = 1, u_{ik} = 0$ results in the same objective value as $r_{Tik} = 1, r_{Tij} = 0, u_{ik} = 1, u_{ij} = 0$. We randomly sample 125 strata from the full match described in Chapter 2. This full match yielded strata of maximal size 8, representing a substantially easier optimization problem than the one presented in Section 3.5.3. The resulting data sets had roughly 500 patients on average. Rather than randomly sampling outcomes, we use the observed outcomes in the randomly sampled matched sets, hence basing this simulation study entirely on real data. In each iteration, we terminated the simulation if either program took longer than 5 minutes to solve in a given iteration. Here, we report total computation time including grouping into unique tables, formulating constants and constraints, and solving the optimization problem.

For formulation (AP1), we found that 29.6% of simulations exceeded the five minute computation limit. Of those that did not, the average computation time was 34.9 seconds for the

pure binary program, but was 0.68 seconds for the linear relaxation. The average relative gap between the optimal binary solution and optimal linear relaxation in the simulations taking under five minutes was 23.5%, representing a marked discrepancy between the linear relaxation and the integer hull of formulation (AP1). Under formulation (P1), all simulations terminated in under five minutes. In fact, the average computation time for our integer program was 0.129 seconds, and the maximal computation time was 0.223 seconds. Among the simulations where alternate formulation (AP1) exceeded our computation time limit, the average computation for our formulation was 0.130 seconds, indicating that our formulation avoids the computational issues due to symmetry that cripple formulation (AP1). The average computation time for the linear relaxation of (P1) was 0.122 seconds. 84.7% of simulated data sets resulted in the optimal integer objective value being equal to that of the linear relaxation. In those iterations where there was a difference, the average relative gap between objective values was a mere 0.003%. Hence, our formulation is markedly stronger than this alternate formulation, as evidenced by reduced computation time even when using the same optimization software: our formulation is over 250 times faster than formulation (AP1) among iterations that solved before computation time ran out, and is thus even faster overall.

B.3.7. Simulation Using Actual Data

In each of 1000 iterations we sample 1250 matched sets from the strata in our motivating example from Section 3.1.2. Each iteration thus has strata ranging in size from 2 to 21, and each data set has an average of roughly 10,000 individuals within it. Large strata affect computation time, as they result in larger numbers of non-exchangeable potential outcome allocations within a stratum and fewer duplicated 2×2 tables in the data. In our data set, 25% of the matched strata had one acute rehabilitation individual and 20 home with home health services patients. Rather than randomly sampling outcomes, we use the observed outcomes in the randomly sampled matched sets, hence basing this simulation study entirely on real data. This simulation setting thus produces particularly challenging optimization

problems: each iteration resulted in over 200,000 variables over which to optimize on average.

We conduct two hypothesis tests in each iteration: a null on the causal risk difference, $\delta = 0.05$, and on the causal risk ratio, $\varphi = 1.10$. For both of the causal estimands being assessed, we test the stated nulls with two-sided alternatives at $\Gamma = 1$ (no unmeasured confounders, integer linear program) and $\Gamma = 1.05$ (unmeasured confounding exists, integer quadratic program). We record the required computation time for each data set, which includes the time for grouping into unique tables, the time taken to define the necessary constants for the problem and also the time required to solve the optimization problem. To measure the strength of our formulation, we also recorded whether or not the initial continuous relaxation had an optimal solution which was itself integral, and if not the relative difference in optimal objective function values between the integer and continuous formulations (defined to be the absolute difference of the two, divided by the absolute value of the relaxed value).

Table 12 shows the results of this simulation study. As one can see, our formulation yields optimal solutions in well under a minute for both the integer linear and integer quadratic formulations despite the magnitude of the problem at hand. The strength of our formulation is further evidenced by the typical discrepancy between the integer optimal solution and that of the continuous relaxation. For testing the causal risk difference, we found that in nearly all of the simulations performed the integer program and its linear relaxation had the *same* optimal objective value. For testing the causal risk ratio, the objective values tended not to be identically equal, which has to do with the existence of fractional values in the row of the constraint matrix enforcing the null hypothesis; nonetheless, the average gap among those iterations where there was a difference was 0.005% percent for the linear program, and 0.01% for the quadratic program. This suggests not only that we have arrived upon a strong formulation, but that one could in practice accurately approximate (P1) by its continuous relaxation.

Table 12: Computation times for tests of $\delta = 0.05$ and $\varphi = 1.10$ at $\Gamma = 1$ (integer linear program) and $\Gamma = 1.05$ (integer quadratic program), along with percentages of coincidence of the integer and relaxed objective values, and average gaps between integer solution and the continuous relaxation if a difference existed between the two.

Null Hypothesis	Avg. Time (s), Integer	Avg. Time (s), Relaxation	$\%(obj_{int} = obj_{rel})$	Avg. Gap If Diff.
$\delta = 0.05; \Gamma = 1.00$	9.26	8.81	99.8%	0.001%
$\delta = 0.05; \Gamma = 1.05$	12.69	8.20	89.5%	0.002%
$\varphi = 1.10; \Gamma = 1.00$	9.74	8.45	9.0%	0.005%
$\varphi = 1.10; \Gamma = 1.05$	13.40	8.38	8.1%	0.011%

B.3.8. Proceeding under Time Constraints

While these simulations suggest that a global integer optimum can be attained in a reasonable amount of time using our formulation, it remains a possibility that for a particular data set the solver may fail to terminate in suitable amount of time for the user. If the user has a maximum allowable period of time for the solver, T_{max}, we would recommend solving the required integer program while terminating the optimization after T_{max} seconds. If the solver terminates before T_{max} then a global integer optimum has been found. Otherwise, integer programming solvers provide bounds on the objective value at any time point t, which can be used to conduct conservative inference and are tighter than those attained by simply solving the continuous relaxation at the outset. Furthermore, one can compare the lower bound to the best integer solution that the solver has found to that point as an indication of how conservative the performed inference truly is.

B.4. Point Estimates for θ Through M-Estimation

While our focus in this work is on inference both assuming and not assuming unmeasured confounding, we briefly describe point estimation for θ. Under the null at $\Gamma = 1$, $T(\theta_0)$ has expectation 0. We propose an m-estimator (also referred to as a z-estimator) for θ by using $T(\theta_0)$ as an estimating function; see Van der Vaart (2000) for more on m- and z- estimators and their corresponding properties. Explicitly, $\hat{\theta} := \mathbf{SOLVE}\{\theta : T(\theta) = 0\}$. This is in keeping with the estimator suggested by Baiocchi et al. (2010) for the effect ratio. For our

three causal estimands of interest, these estimators are:

$$\hat{\delta} = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{n_i} n_i \left(Z_{ij}R_{ij}/m_i - (1-Z_{ij})R_{ij}/(n_i - m_i)\right)$$

$$\hat{\varphi} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} n_i Z_{ij}\frac{R_{ij}}{m_i}}{\sum_{i=1}^{I}\sum_{j=1}^{n_i} n_i (1-Z_{ij})\frac{R_{ij}}{n_i - m_i}}$$

$$\hat{\lambda} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n_i} n_i \left(Z_{ij}\frac{R_{ij}}{m_i} - (1-Z_i)\frac{R_{ij}}{n_i - m_i}\right)}{\sum_{i=1}^{I}\sum_{j=1}^{n_i} n_i \left(Z_{ij}\frac{D_{ij}}{m_i} - (1-Z_i)\frac{D_{ij}}{n_i - m_i}\right)}.$$

While useful as indications of effect magnitude size, these estimators do not play a direct role in conducting inference or performing sensitivity analyses; rather, our focus lies in understanding the randomization distribution of $T(\theta_0)$ at any particular value of θ_0. Confidence intervals under no unmeasured confounding are then constructed by inverting tests for a sequence of null hypotheses. Constructing intervals in this manner avoids certain issues associated with intervals directly based on m-estimators, such as small sample bias and heavy dependence of the estimator's variance on the estimand of interest; see Chapter 2 for a discussion of the latter point as it pertains to constructing confidence intervals for the risk difference within a matched observational study.

B.5. Assuming a Known Direction of Effect Impacts Reported Sensitivity

In both examples in Section 3.6, we perform inference under a host of assumptions on the potential outcomes. As is demonstrated therein, the assumption of a known direction of effect has a particularly strong impact on the corresponding sensitivity analysis. Note that when testing the null of $\delta = 0 \Leftrightarrow \varphi = 1 \Leftrightarrow \lambda = 0$ under the assumption of a direction of effect, the *only* allocation of $\mathbf{r}_T, \mathbf{r}_C$ that satisfies the null hypothesis is the allocation of Fisher's sharp null: $r_{Tij} = r_{Cij}\ \forall i, j$. This results in testing a simple, rather than composite, null hypothesis. At $\Gamma = 1$, the necessary hypothesis test can be performed using the permutation distribution (or a normal approximation thereof) of the test statistic under Fisher's sharp null. For $\Gamma > 1$ the potential outcomes are still fixed at those of Fisher's sharp null, but

we must consider the possible vectors of unmeasured confounders. Without the assumption of a direction of effect, there are many possible allocations of potential outcomes satisfying this null. This additional flexibility in the optimization problem results in more extreme worst-case allocations for the inference being conducted.

As a simple illustration of why this is the case, consider testing this null with two pairs of individuals. In stratum 1, suppose $R_{11} = R_{12} = 1$, while in stratum 2 suppose $R_{21} = R_{22} = 0$, where without loss of generality the first individual in each matched set received the treatment. If we assume a nonnegative treatment effect, $r_{T12} = 1$, since $r_{C12} = 1$. Similarly, $r_{C21} = 0$ since $r_{T21} = 0$. Finally, the constraint that the null is true forces $r_{C11} = 1$ and $r_{T22} = 0$. For any Γ, these strata contribute expectation and variance 0. Without the assumption of a direction of effect, we can also satisfy the null hypothesis by setting $r_{C11} = 1$, $r_{T12} = 0$, $r_{C21} = 0$, $r_{T22} = 1$. Not only would we then have positive variance contribution from each of these strata at any Γ, but also setting $\mathbf{u}_1 = [1, 0]$ and $\mathbf{u}_2 = [0, 1]$ results in an aggregate expected value of $(\Gamma - 1)/(1 + \Gamma) \geq 0$. These choices allow one to find a less significant deviate under no constraints on the direction of effect than is possible under a model with a known direction of effect.

B.6. Sensitivity Analysis for a Simple Null

While the methodology presented herein was motivated by conducting sensitivity analyses for composite null hypotheses with binary outcomes, we note that a simplified version can be used to conduct a sensitivity analysis for a simple null hypothesis for general types of outcome variables without invoking asymptotic separability (Gastwirth et al., 2000). With a simple null hypothesis, q_{ij} are fixed for each individual i and each stratum j. In the notation of Section 3.4, S represents the number of strata with unique sets of values for the vector $\mathbf{q}_i$. With continuous outcomes S will often equal I, but for other types of outcomes there may be repeated strata. For each s, P_s (the number of possible allocations of potential outcomes within unique set s) equals 1 as both sets of potential outcomes are fixed under a simple null. Hence, the subscript $[sp]$ in our original formulation can be replaced by a single

subscript s, Define μ_{sa} and ν_{sa} by replacing $[sp]$ with s in the notation of Section 3.4.1, and make the analogous substitution of x_{sa} for $x_{[sp]a}$. Let M_s again represent the number of times unique stratum s occurred, and let $\tilde{n}_s$ be the number of observations within unique stratum s. Define $\boldsymbol{\mu} = [\mu_{11}, .., \mu_{S,\tilde{n}_S-1}]$ and let the analogous definitions hold for $\boldsymbol{\nu}$ and $\mathbf{x}$. Finally, note that the constraint that the null must be true in formulation (P1) can be removed entirely as q_{ij} are defined under this assumption. A sensitivity analysis at a given $\Gamma > 1$ can be conducted by solving the following optimization problem:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \;\; & (t - (\boldsymbol{\mu}^T\mathbf{x}))^2 - \kappa(\boldsymbol{\nu}^T\mathbf{x}) \\ \text{subject to} \;\; & \sum_{a=1}^{\tilde{n}_s-1} x_{sa} = M_s \;\; \forall s \\ & x_{sa} \in \mathbb{Z} \;\; \forall s, a \\ & x_{sa} \geq 0 \;\; \forall s, a \end{aligned} \tag{P2}$$

As described in Section 3.5.2, we can conduct a sensitivity analysis for a given $\Gamma > 1$ by minimizing (P2) with $\kappa = \chi^2_{1,1-\alpha}$. To find the actual minimal deviate, we can follow the iterative procedure outlined in Section 3.5.2 until converging to a stationary κ^*.

The constraint matrix corresponding to the above optimization program is *totally unimodular*. As a consequence, the polyhedron of the continuous relaxation equals the integer hull (Bertsimas and Tsitsiklis, 1997). Hence, if one were solving an integer *linear* program, the solution of the continuous relaxation would be guaranteed to be integral. When finding the worst-case deviate we are minimizing a constrained convex quadratic function; as such, the solution need not be at the vertex. Nonetheless, strong formulations of integer quadratic programs are essential for efficiently finding optimal solutions.

B.6.1. Example: Dropping Out of High School and Cognitive Achievement

As an exposition of their methodology, Gastwirth et al. (2000) consider conducting a sensitivity analysis for comparing cognitive achievement of US high-school drop-outs with that

of non-dropouts; see Rosenbaum (1986) for more details on the study. They conducted inference on 12 drop-outs in the study, where each drop-out was matched to two students who did not drop out, yet were similar on the basis of all other observed covariates. Using an aligned rank test, the test statistic for these 12 matched sets was $t = 296$, with expectation and variance at $\Gamma = 1$ of 222 and 1271, yielding a standardized deviate of 2.07 and approximate one sided p-value of 0.019.

Table 3 of Gastwirth et al. (2000) shows the results of the asymptotically separable algorithm on this data set for $\Gamma = 2$. At this strength of unmeasured confounding, the separable algorithm yields a bounding normal deviate with a mean of 257.40 and a variance of 1177.23, resulting in an approximation to the worst-case deviate of 1.125 and a one sided p-value of 0.129. We also explicitly minimized the deviate by solving (P2). This yields a bounding random variable with a mean of 256.60 and a variance of 1228.145, yielding a worst-case deviate of 1.124 and a worst-case p-value of 0.130. Investigating further, the worst-case allocations of $\mathbf{u}$ for each stratum were in agreement for all of the matched sets except for matched set 11. There, the asymptotically separable algorithm chooses $\mathbf{u}_{11} = [0, 1, 1]$, contributing a mean of 24.80 and a variance of 139.76. The correct value for $\mathbf{u}_{11}$ for minimizing the deviate is $\mathbf{u}_{11} = [0, 0, 1]$, which has slightly lower expectation (24.24) but larger variance (173.19).

This demonstrates that for I even moderately large, the asymptotically separable algorithm can produce a bounding random variable that very closely approximates the true upper bound on the p-value. That being said, given our formulation the worst-case deviate can be explicitly found. Furthermore, one need not worry about computation time: for conducting the sensitivity analysis on this problem, an optimal solution was found in 0.15 seconds.

APPENDIX C

C.1. Additional Details for the Smoking and Naphthalene Example

Following Weitzman et al. (2005) and Suwan-ampai et al. (2009), individuals were classified as active smokers if they stated that they smoke "every day" or "some days" in response to the question "Do you now smoke cigarettes?," or if their serum cotinine (a metabolite of nicotine) levels were above 0.05 ng/mL. Using this definition, there were 453 smokers and 1253 nonsmokers. The nonsmokers include former smokers and never smokers, as urinary naphthol is an indicator of recent naphthalene exposure.

We used full matching to control for observed covariates in this study (Rosenbaum, 1991; Hansen, 2004). In this match, we allowed for strata of maximal size 10, meaning that a matched set could have, at most, either 1 current smoker and 9 nonsmokers; or 1 nonsmoker and 9 current smokers. We identified 22 pre-treatment covariates deemed predictive of smoking and naphthalene levels based on those used in Suwan-ampai et al. (2009); these covariates are listed in Figure 15. Ten covariates contained missing values, with a maximal percentage of values missing of 10%. To account for this, we included 10 missingness indicators as additional covariates upon which to match. As discussed in Rosenbaum and Rubin (1984) and Rosenbaum (2010, Section 9.4), this facilitates balancing the observed covariates and the pattern of missingness. Rank-based Mahalanobis distance with a propensity score caliper of 0.08 was used, and propensity scores were estimated using logistic regression (Rosenbaum, 2010, Section 8.3). Figure 15 shows the standardized differences before and after matching for observed confounders and demonstrates that before matching there were substantial imbalances between smokers and nonsmokers with respect to many important variables. It also shows that matching was able to effectively create a well-balanced comparison between smokers and nonsmokers on the basis of these variables. Details for calculating standardized differences before and after full matching can be found in Stuart and Green (2008) and Rosenbaum (2010, Section 9.1).

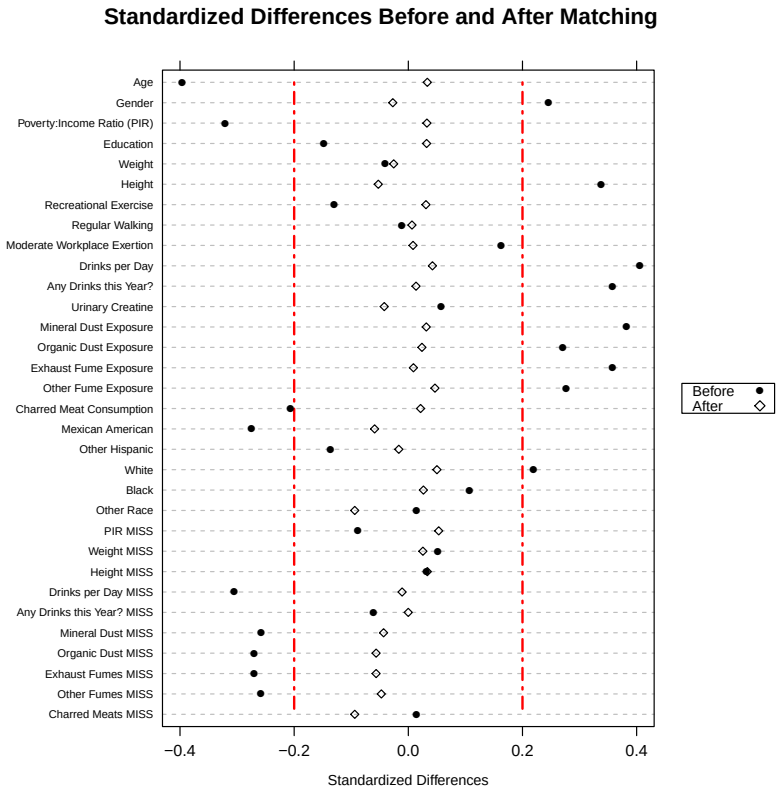

Figure 15: Covariate Imbalances Before and After Matching. The dotplot (a Love plot) shows the absolute standardized differences without matching, and after conducting a matching with a variable number of controls. The vertical dotted line corresponds to a standardized difference threshold of 0.2, which is often regarded as the maximal allowable absolute standardized difference (for example, Silber et al., 2001). The largest absolute standardized difference after matching was 0.094.

C.2. A Simple Extension To One-Sided Testing

By taking the square of the deviate in our original formulation, we lose the deviate's sign. While this does not make a difference for two-sided testing, it does make a difference when the test is one-sided. For example, if we stipulated a one-sided, greater than alternative but observed a test statistic markedly *smaller* than its expectation under the null we should fail to reject that null, a fact which is lost when taking the square. To account for this, we introduce a penalty into the constraints corresponding to one-sided hypotheses that only allow for a rejection to be registered if the expectation of the test statistic yielded through the sensitivity analysis maintains the proper relationship with the observed test statistic

given the nature of the alternative. Let b_k be a binary variable for the k^{th} outcome, and let M be a sufficiently large constant.

Redefine $\zeta_k(\boldsymbol{\varrho})$ so that

$$\zeta_k(\boldsymbol{\varrho}) = \begin{cases} (t_k - \mathbb{E}[t_k(\mathbf{Z}, \mathbf{F}_k); \boldsymbol{\varrho}])^2 - \chi^2_{1,1-\alpha/K}\mathbf{Var}(t_k(\mathbf{Z}, \mathbf{F}_k); \boldsymbol{\varrho}) & \text{if two-sided alternative} \\ (t_k - \mathbb{E}[t_k(\mathbf{Z}, \mathbf{F}_k); \boldsymbol{\varrho}])^2 - \chi^2_{1,1-2\alpha/K}\mathbf{Var}(t_k(\mathbf{Z}, \mathbf{F}_k); \boldsymbol{\varrho}) & \text{if one-sided alternative} \end{cases}$$

We then modify our optimization problem as follows:

$$\begin{aligned} &\underset{y,\varrho_{ij},s_i,b_k}{\text{minimize}} \ y \\ &\text{subject to} \ \ y \geq \zeta_k(\boldsymbol{\varrho}) - Mb_k \ \ \forall k \\ &\qquad \sum_{j=1}^{n_i} \varrho_{ij} = 1 \ \ \forall i \\ &\qquad s_i \leq \varrho_{ij} \leq \Gamma s_i \ \ \forall i, j \\ &\qquad \varrho_{ij} \geq 0 \ \ \forall i, j \\ &\qquad b_k \in \{0, 1\} \ \ \forall k \\ &\qquad b_k = 0 \ \text{ if } H_k \text{ two-sided} \\ -M(1 - b_k) \leq t_k - \boldsymbol{\varrho}^T\mathbf{q}_k \leq Mb_k \ &\text{ if } H_k \text{ one-sided }, < \\ -Mb_k \leq t_k - \boldsymbol{\varrho}^T\mathbf{q}_k \leq M(1 - b_k) \ &\text{ if } H_k \text{ one-sided }, > \end{aligned}$$

The value Mb_k added to the k constraints on the auxiliary variable y, in conjunction with the constraints on the value of the test statistic's numerator, impose a heavy negative penalty if the relationship between the test statistic and its mean under a given allocation of unmea-

Table 13: Rejection probability for testing true and false nulls through closed testing. Desired strong familywise error control at 0.05.

		True Nulls			False Nulls	
Gamma	Moments	H_1	H_2	$H_1 \wedge H_2$	H_3	$H_1 \wedge H_2 \wedge H_3$
$\Gamma = 1$	$\boldsymbol{\tau}, \Sigma^{(1)}$	0.0260	0.0266	0.0506	0.9884	0.9886
	$\boldsymbol{\tau}, \Sigma^{(2)}$	0.0267	0.0268	0.0462	0.9881	0.9893
$\Gamma = 1.05$	$\boldsymbol{\tau}, \Sigma^{(1)}$	0.0102	0.0089	0.0189	0.9748	0.9749
	$\boldsymbol{\tau}, \Sigma^{(2)}$	0.0096	0.0122	0.0197	0.9732	0.9750
$\Gamma = 1.10$	$\boldsymbol{\tau}, \Sigma^{(1)}$	0.0035	0.0043	0.0078	0.9462	0.9463
	$\boldsymbol{\tau}, \Sigma^{(2)}$	0.0053	0.0032	0.0081	0.9441	0.9462

sured confounders do not adhere to the required direction imposed by the alternative. This makes it such that we will never reject a null at a given Γ because a given one-sided test was highly *insignificant*, which without such a penalty would be construed as being highly significant.

C.3. Simulation of Type I Error Control

In this simulation study, we demonstrate that, in the presence of true intersection null hypotheses, our procedure strongly controls the familywise error rate at level $\alpha = 0.05$. In each of 6 simulation settings, we simulate 10,000 data sets under no unmeasured confounding with $I = 250$ pairs for three outcome variables of interest and using Huber's M-statistic, as described in Section 4.6. For each of the 2 combinations of treatment effects and covariances, closed testing is used, with our minimax procedure being used for each intersection null. Tests are run at $\Gamma = 1, 1.05$, and 1.1. The values for the treatment effect vector and the covariances were as follows:

1. $\boldsymbol{\tau} = [0, 0, 0.3]$

2. $\Sigma^{(1)} = \text{Diag}(1)$; $\Sigma^{(2)}_{ij} = 1$ if $i = j$, $\Sigma^{(2)}_{ij} = 0.5$ otherwise.

We test Fisher's sharp null on each outcome. In each iteration, we record whether or not the true null hypotheses H_1, H_2, and $H_1 \wedge H_2$ are rejected. We also record whether or not the false nulls H_3 and $H_1 \wedge H_2 \wedge H_3$ are rejected. Table 13 shows the results of this simulation

study. As can be seen, our procedure strongly controls the type I error rate for all values of Γ tested. The rate of rejection for $H_1 \wedge H_2$ reveals that our procedure is conservative when the test statistics are dependent, while coming very close to attaining the actually desired level under independence. As Γ increases the Type I error rate decreases for all true nulls, as many spurious rejections assuming no unmeasured confounding can be explained by moderate departures from pure randomization.

APPENDIX D

D.1. Proof of Theorem 2

We begin by restating the optimization problem

$$\begin{aligned}
\underset{\{\pi_i,\varphi_{i2}\}}{\text{minimize}} \quad & \frac{\sum_{i=1}^{I}\varphi_{i1}-\sum_{i=1}^{I}\left(\pi_i\varphi_{i1}+(1-\pi_i)\varphi_{i2}\right)}{\sqrt{\frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I}(\varphi_{i1}-\varphi_{i2})^2}} \qquad \text{(P2)}\\
\text{subject to} \quad & \sum_{i=1}^{I}\varphi_{i1}+\varphi_{i2}=2I\bar{\Delta}_0\\
& \frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I}(\varphi_{i1}-\varphi_{i2})^2=\bar{\bar{v}}_\Gamma\\
& \pi_i=w_i/(1+\Gamma)+(1-w_i)\Gamma/(1+\Gamma)\\
& w_i=\mathbb{1}\{\varphi_{i2}\geq\varphi_{i1}\}
\end{aligned}$$

Before proceeding, we prove two useful lemmas:

Lemma 3. *For $\sum_{i=1}^{I} w_i \in \{1,...,I-2\}$, any feasible solution to Problem (P2) involving $\varphi_{i2}=\varphi_{i1}$ for some i has an objective value that is greater than or equal to a feasible solution with $\varphi_{i2}\neq\varphi_{i1}\forall i$.*

Proof. Suppose $\varphi_{i1}=\varphi_{i2}$ and $\sum_{i=1}^{I} w_i \in \{1,...,I-2\}$. Then, there exist two pairs, j and k, such that $\varphi_{j2}-\varphi_{j1}<0$ and $\varphi_{k2}-\varphi_{k1}<0$. Define $\tilde{\varphi}_{i2}=\varphi_{i2}-c$, $\tilde{\varphi}_{j2}=\varphi_{j2}+c/2$, and $\tilde{\varphi}_{k2}=\varphi_{k2}+c/2$.

First, note that the change to the numerator of the objective function is less than or equal to as changing to $\tilde{\varphi}_{i2}$ decreases by $c/(1+\Gamma)$, while changing to $\tilde{\varphi}_{j2}$ and $\tilde{\varphi}_{k2}$ increases it by $(c/2+c/2)/(1+\Gamma)$ if $\tilde{\varphi}_{j2}<\tilde{\varphi}_{j1}$ and $\tilde{\varphi}_{k2}<\tilde{\varphi}_{k1}$. If one of these inequalities reverses based on the value of c, the change in numerator will be negative.

We now evaluate the impact on the variance. Computing: $(\tilde{\varphi}_{i2} - \varphi_{i1})^2 + (\tilde{\varphi}_{j2} - \varphi_{j1})^2 + (\tilde{\varphi}_{k2} - \varphi_{k1})^2 = (\varphi_{i2} - \varphi_{i1})^2 + (\varphi_{j2} - \varphi_{j1})^2 + (\varphi_{k2} - \varphi_{k1})^2 + 3c^2/2 + c(\varphi_{j2} - \varphi_{j1} + \varphi_{k2} - \varphi_{k1})$. Setting $c = (2/3)(\varphi_{j1} - \varphi_{j2} + \varphi_{k1} - \varphi_{k2}) > 0$ yields the identical $\square$

Hence we can assume that $\sum_{i=1}^{I} w_i \in \{1, .., I-2\} \Rightarrow (w_i = 1 \Rightarrow \{\varphi_{i2} < \varphi_{i1}\})$. That is, $\sum_{i=1}^{I} w_i \in \{1, .., I-2\} \Rightarrow \varphi_{i1} \neq \varphi_{i2} \ \ \forall i$.

Lemma 4. *Suppose* $\sum_{i=1}^{I} w_i in \{1, ..., I-2\}$. *Then, at the solution to the problem above,* $\frac{\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{I} (\varphi_{i1} - \varphi_{i2})^2 = \bar{\bar{v}}_{\Gamma}$

Proof. For any feasible set of $\{\varphi_{i2}\}$, by Lemma 3 we can find φ_{i2} and $\varphi_{i'2}$ such that $\varphi_{i2} > \varphi_{i1}$ and $\varphi_{i'2} < \varphi_{i'1}$. Define $\tilde{\varphi}_{i2} = \varphi_{i2} + c$ and $\tilde{\varphi}_{i'2} = \varphi_{i'2} - c$ with $c > 0$. Replacing φ_{i2} and $\varphi_{i'2}$ with $\tilde{\varphi}_{i2}$ and $\tilde{\varphi}_{i'2}$, the constraint imposed by the null is still satisfied. Furthermore, the numerator of the objective function changes by $-(\Gamma - 1)/(1 + \Gamma)c$, while the denominator increases by $\Gamma/(1+\Gamma)^2(2c^2 + 2c(\varphi_{i2} - \varphi_{i1} + \varphi_{i'1} - \varphi_{i'2})) > 0$. The objective function is thus further minimized, and c can be chosen such that the variance constraint is still satisfied. $\square$

Proof of Theorem 2. Using Lemmas 3-4, we can, for $\sum_{i=1}^{I} w_i \in \{1, ..., I-2\}$, reformulate the optimization problem as one which seeks to maximize the expectation of the average treatment effect

$$
\begin{aligned}
\underset{\{\varphi_{i2}\}}{\text{maximize}} \quad & \sum_{i=1}^{I} (\pi_i \varphi_{i1} + (1 - \pi_i)\varphi_{i2}) \\
\text{subject to} \quad & \sum_{i=1}^{I} \varphi_{i1} + \varphi_{i2} = 2I\bar{\Delta}_0 \\
& \frac{\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{I} (\varphi_{i1} - \varphi_{i2})^2 = \bar{\bar{v}}_{\Gamma} \\
& \pi_i = w_i/(1+\Gamma) + (1 - w_i)\Gamma/(1+\Gamma) \\
& w_i = \mathbb{1}\{\varphi_{i2} \geq \varphi_{i1}\}
\end{aligned}
$$

The Lagrangian of the above problem is :

$$\mathcal{L} = \sum_{i=1}^{I} (\pi_i \varphi_{i1} + (1-\pi_i)\varphi_{i2}) + \lambda_1 \left(\sum_{i=1}^{I} \varphi_{i1} + \varphi_{i2} - 2I\bar{\Delta}_0 \right) + \lambda_2 \left(\sum_{i=1}^{I} (\varphi_{i1} - \varphi_{i2})^2 - \frac{(1+\Gamma)^2}{\Gamma} \bar{\bar{v}}_\Gamma \right)$$

Differentiating with respect to φ_{i2} and setting to zero yields:

$$0 = w_i \Gamma/(1+\Gamma) + (1-w_i)/(1+\Gamma) + \lambda_1 + 2\lambda_2(\varphi_{i2} - \varphi_{i1})$$

This form then implies

$$\varphi_{i2} - \varphi_{i1} = \begin{cases} \frac{-\Gamma/(1+\Gamma)-\lambda_1}{2\lambda_2} & \varphi_{i2} \geq \varphi_{i1} \\ \frac{-1/(1+\Gamma)-\lambda_1}{2\lambda_2} & \varphi_{i2} < \varphi_{i1} \end{cases}$$

By Lemma 3, it must be the case that $C^+ := \frac{-\Gamma/(1+\Gamma)-\lambda_1}{2\lambda_2} > 0$ and $C^- := \frac{-1/(1+\Gamma)-\lambda_1}{2\lambda_2} < 0$.

Hence, we can now further simplify the form of the optimization problem:

$$
\begin{aligned}
\underset{\{C^+,C^-,w_i\}}{\text{maximize}} \quad & \sum_{i=1}^{I} \varphi_{i1} + \frac{\Gamma C^+}{1+\Gamma} \sum_{i=1}^{I} w_i + \frac{C^-}{1+\Gamma} \sum_{i=1}^{I} (1-w_i) \\
\text{subject to} \quad & \sum_{i=1}^{I} 2\varphi_{i1} + C^- \sum_{i=1}^{I} (1-w_i) + C^+ \sum_{i=1}^{I} w_i = 2I\bar{\Delta}_0 \\
& \frac{\Gamma}{(1+\Gamma)^2} \sum_{i=1}^{I} \left((C^-)^2 (1-w_i) + (C^+)^2 w_i \right) = \bar{\bar{v}}_\Gamma \\
& C^+ \geq 0 \\
& C^- \leq 0 \\
& \sum_{i=1}^{I} w_i \in \{1, ..., I-2\}
\end{aligned}
$$

For $\sum_{i=1}^{I} w_i \in \{1, ..., I-2\}$, we can express the optimal C^+ and C^- as functions of $s = \sum_{i=1}^{I} w_i$ as:

$$
C_s^+ = \frac{2I\bar{\Delta}_0 - 2\sum_{i=1}^{I} \varphi_{i1} - C_s^-(I-s)}{s}
$$

$$
C_s^- = \frac{4\sum_{i=1}^{I} (\bar{\Delta}_0 - \varphi_{i1}) \frac{I-s}{s} - 2\sqrt{\left(\frac{I-s}{s}\right) \left(I \frac{(1+\Gamma)^2}{\Gamma} \bar{\bar{v}}_\Gamma - 4 \left(\sum_{i=1}^{I} (\bar{\Delta}_0 - \varphi_{i1}) \right)^2 \right)}}{2I \left(\frac{I-s}{s} \right)}
$$

The identity for C_s^+ follows trivially from the constraint imposed by the null, $\sum_{i=1}^{I} \varphi_{i1} +$

$\varphi_{i2} = 2I\bar{\Delta}_0$:

$$\sum_{i=1}^{I} \varphi_{i1} + \varphi_{i2} = 2I\bar{\Delta}_0$$

$$\sum_{i=1}^{I} 2\varphi_{i1} + w_i C_s^+ + (1 - w_i)C_s^- = 2I\bar{\Delta}_0$$

$$C_s^+ \sum_{i=1}^{I} w_i + C_s^- \sum_{i=1}^{I}(1 - w_i) = 2I\bar{\Delta}_0 - 2\sum_{i=1}^{I} \varphi_{i1}$$

$$\frac{2I\bar{\Delta}_0 - 2\sum_{i=1}^{I} \varphi_{i1} - C_s^- \sum_{i=1}^{I}(1 - w_i)}{\sum_{i=1}^{I} w_i} = C_s^+$$

$$\frac{2I\bar{\Delta}_0 - 2\sum_{i=1}^{I} \varphi_{i1} - C_s^-(I - s)}{s} = C_s^+$$

To derive the expression for C_s^+, note that the variance equality implies:

$$(C_s^+)^2 s + (C_s^-)^2(I - s) = \bar{\bar{v}}\frac{(1 + \Gamma)^2}{\Gamma}$$

Expressing C_s^+ in terms of C_s^- and using the quadratic formula then yields the expression. These values of C^+ then yield the values for μ_s for $s \in \{1, ..., I - 2\}$ given in the Section 5.4.1., and $\nu_s^2 = \bar{\bar{v}}_\Gamma$ for $s \in \{1, ..., I - 2\}$ by Lemma 4.

We now consider the set of solutions for which $\sum_{i=1}^{I} w_i \in \{0, I - 1, I\}$.

(1) $\sum_{i=1}^{I} w_i = I$. Given $\bar{y} \geq \bar{\Delta}_0$ as we have assumed, $\sum_{i=1}^{I} w_i = I$ cannot be a solution, as in this case the constraint imposed by the null cannot be satisfied.

(2) $\sum_{i=1}^{I} w_i = 0$ At $\sum_{i=1}^{I} w_i = 0$, the numerator of the objective value will then be fixed, due to the constraint imposed by the null, at $2/(1 + \Gamma)\sum_{i=1}^{I}(\varphi_{i1} - \bar{\Delta}_0)$. Thus, minimizing the objective function is achieved by maximizing its denominator:

$$\begin{aligned}
\underset{\{\varphi_{i2}\}}{\text{maximize}} \quad & \frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I}(\varphi_{i1}-\varphi_{i2})^2 \\
\text{subject to} \quad & \sum_{i=1}^{I}\varphi_{i1}+\varphi_{i2}=2I\bar{\Delta}_0 \\
& \frac{\Gamma}{(1+\Gamma)^2}\sum_{i=1}^{I}(\varphi_{i1}-\varphi_{i2})^2 \leq \bar{\bar{v}}_\Gamma
\end{aligned}$$

The KKT conditions, along with $\sum_{i=1}^{I} w_i = 0$, require the following to hold:

$$2(\varphi_{i2}-\varphi_{i1}) = \lambda_1 + 2\lambda_2(\varphi_{i2}-\varphi_{i1})$$

For all i, given $\sum_{i=1}^{I} w_i = 0$ we know $\varphi_{i2} < \varphi_{i1}$. Based on the KKT condition, it is clear that $\varphi_{i2} = \varphi_{i1} + C^-$ for some constant $C^- < 0$, as rearranging the above yields $(\varphi_{i2}-\varphi_{i1}) = \lambda_1/(2-2\lambda_2)\ \forall i$.

With this in mind, from the constraint imposed by the null we have that $C^- = 2I^{-1}\sum_{i=1}^{I}(\bar{\Delta}_0 - \varphi_{i1})$, yielding an objective value of $\frac{\Gamma}{(1+\Gamma)^2}4I^{-1}(\sum_{i=1}^{I}(\bar{\Delta}_0-\varphi_{i1}))^2$. If this satisfies the variance inequality, it is a valid solution; otherwise, there is no potentially optimal solution with $\sum_{i=1}^{I} w_i = 0$. This leads to the values for μ_0 and ν_0 given in the Section 5.4.1.

(3) $\sum_{i=1}^{I} w_i = I - 1$,

In this scenario, it could be the case that $\varphi_{i1} = \varphi_{i2}$ for $I-1$ of I pairs, while $\varphi_{i'2} - \varphi_{i'1} = \sum_{i=1}^{I}(2\bar{\Delta_0} - \varphi_{i1})$. If it so happens that $\frac{\Gamma}{(1+\Gamma)^2}\left(\sum_{i=1}^{I}(2\bar{\Delta_0}-\varphi_{i1})\right)^2 = \bar{\bar{v}}_\Gamma$, this is a feasible solution; however, it cannot be optimal, as is now shown.

Take $j,k \neq i'$. such that $\varphi_{j2}-\varphi_{j1} = 0$ and $\varphi_{k2}-\varphi_{k1} = 0$. Define $\tilde{\varphi}_{i'2} = \varphi_{i'2} + c$,

$\tilde{\varphi}_{j2} = \varphi_{j2} - c/2$, and $\tilde{\varphi}_{k2} = \varphi_{k2} - c/2$.

Evaluating further, $(\tilde{\varphi}_{i2} - \varphi_{i1})^2 + (\tilde{\varphi}_{j2} - \varphi_{j1})^2 + (\tilde{\varphi}_{k2} - \varphi_{k1})^2 = (\varphi_{i2} - \varphi_{i1})^2 + 3c^2/2 + 2c((\varphi_{i'2} - \varphi_{i'1})$. Setting $c = (4/3)(\varphi_{i'1} - \varphi_{j2}) > 0$ yields an objective value that is improved by $(1/3)\Gamma/(1+\Gamma)(\varphi_{i'1} - \varphi_{j2})$ and now has $\sum_{i=1}^{I} w_i = I - 2$, thus putting us in the regime previously considered.

If the variance constraint is not binding, the objective value can be improved by changing the values of φ_{i2} such that *none* of them exactly φ_{i1}. If this is the case, the solution when $\sum_{i=1}^{I} w_i = I - 1$ can be attained as it was when $\sum_{i=1}^{I} w_i \in \{1, ..., I-2\}$, thus recovering the values for μ_{I-1} and ν^2_{I-1} given in Section 5.4.1.

BIBLIOGRAPHY

J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

D. C. Angus and T. van der Poll. Severe sepsis and septic shock. *New England Journal of Medicine*, 369(9):840–851, 2013.

D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pinsky. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Critical Care Medicine*, 29(7):1303–1310, 2001.

P. M. Aronow, D. P. Green, and D. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42(3):850–871, 2014.

M. Baiocchi, D. S. Small, S. Lorch, and P. R. Rosenbaum. Building a stronger instrument in an observational study of perinatal care for premature infants. *Journal of the American Statistical Association*, 105(492):1285–1296, 2010.

D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *Annals of Statistics*, 44(2):813–852, 2016.

C. Brun-Buisson, F. Doyon, and J. Carlet. Bacteremia and severe sepsis in adults: a multicenter prospective survey in ICUs and wards of 24 hospitals. *American Journal of Respiratory and Critical Care Medicine*, 154(3):617–624, 1996.

D. Campbell. Definitional vs multiple operationalism. In E. Overman, editor, *Methodology and Epistemology for Social Science*, pages 32–36. University of Chicago Press, Chicago, 1988.

C. Charalambous and A. R. Conn. An efficient method to solve the minimax problem directly. *SIAM Journal on Numerical Analysis*, 15(1):162–187, 1978. ISSN 00361429.

A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Research Logistics Quarterly*, 9(3-4):181–186, 1962.

H. A. Chipman, E. I. George, and R. E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010.

W. G. Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128(2):234–266, 1965.

R. J. Cook and D. L. Sackett. The number needed to treat: a clinically useful measure of treatment effect. *British Medical Journal*, 310(6977):452–454, 1995.

J. Cornfield, W. Haenszel, E. C. Hammond, A. M. Lilienfeld, M. B. Shimkin, and E. L. Wynder. Smoking and lung cancer:recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22:173–203, 1959.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009.

R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448): 1053–1062, 1999.

P. Ding. A paradox from randomization-based causal inference. *arXiv preprint arXiv:1402.0142*, 2014.

P. Ding and L. W. Miratrix. To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. *Journal of Causal Inference*, (3):41–57, 2014.

P. Ding and T. J. Vanderweele. Generalized Cornfield conditions for the risk difference. *Biometrika*, 101(4):971–977, 2014.

W. Dinkelbach. On nonlinear fractional programming. *Management Science*, 13(7):492–498, 1967.

H. F. Dorn. Philosophy of inferences from retrospective studies. *American Journal of Public Health and the Nations Health*, 43(6):677–683, 1953.

J. Eckstein, P. L. Hammer, Y. Liu, M. Nediak, and B. Simeone. The maximum box problem and its application to data analysis. *Computational Optimization and Applications*, 23 (3):285–298, 2002.

B. L. Egleston, D. O. Scharfstein, and E. MacKenzie. On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics*, 65(2):497–504, 2009.

A. Esteban, F. Frutos-Vivar, N. D. Ferguson, O. Peñuelas, J. Á. Lorente, F. Gordo, T. Honrubia, A. Algora, A. Bustos, and G. García. Sepsis incidence and outcome: Contrasting the intensive care unit with the hospital ward. *Critical Care Medicine*, 35(5):1284–1289, 2007.

H. Finner and K. Strassburger. The partitioning principle: a powerful tool in multiple decision theory. *Annals of Statistics*, 30(4):1194–1213, 2002.

R. A. Fisher. *The Design of Experiments*. Oliver & Boyd, 1935.

L. Forrow, W. C. Taylor, and R. M. Arnold. Absolutely relative: how research results are summarized can affect treatment decisions. *The American Journal of Medicine*, 92(2): 121–124, 1992.

D. F. Gaieski, J. M. Edwards, M. J. Kallan, and B. G. Carr. Benchmarking the incidence and mortality of severe sepsis in the United States. *Critical Care Medicine*, 41(5):1167–1174, 2013.

J. L. Gastwirth, A. M. Krieger, and P. R. Rosenbaum. Asymptotic separability in sensitivity analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62 (3):545–555, 2000.

F. Glover and E. Woolsey. Converting the 0-1 polynomial programming problem to a 0-1 linear program. *Operations Research*, 22(1):180–182, 1974.

J. J. Goeman and L. Finos. The inheritance procedure: multiple testing of tree-structured hypotheses. *Statistical Applications in Genetics and Molecular Biology*, 11(1):1–18, 2012.

J. J. Goeman and A. Solari. The sequential rejection principle of familywise error control. *Annals of Statistics*, 38(6):3782–3810, 2010.

S. Greenland. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306, 2003.

S. Greenland and J. M. Robins. Estimation of a common effect parameter from sparse follow-up data. *Biometrics*, 41(1):55–68, 1985.

A. P. Grieve. The number needed to treat: a useful clinical measure or a case of the emperor's new clothes? *Pharmaceutical Statistics*, 2(2):87–102, 2003.

J. Hájek and Z. Šidák. Theory of Rank Tests. 1967.

E. C. Hammond. Smoking in relation to mortality and morbidity. findings in first thirty-four months of follow-up in a prospective study started in 1959. *Journal of the National Cancer Institute*, 32(5):1161–1188, 1964.

B. B. Hansen. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618, 2004.

B. B. Hansen and S. O. Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 2012.

S. S. Hecht. Human urinary carcinogen metabolites: biomarkers for investigating tobacco and cancer. *Carcinogenesis*, 23(6):907–922, 2002.

J. J. Heckman, S. Urzua, and E. Vytlacil. Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, 88(3):389–432, 2006.

M. A. Hernán and J. M. Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, 2016.

J. Hill and Y. Su. Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, 7(3):1386–1420, 2013.

J. Hill, C. Weiss, and F. Zhai. Challenges with propensity score strategies in a high-dimensional setting and a potential alternative. *Multivariate Behavioral Research*, 46 (3):477–513, 2011.

J. Hodges and E. L. Lehmann. Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics*, 33(2):482–497, 1962.

J. L. Hodges and E. L. Lehmann. Estimates of location based on rank tests. *The Annals of Mathematical Statistics*, 34(2):598–611, 1963.

P. W. Holland. Causal inference, path analysis and recursive structural equations models. *Sociological Methodology*, 18:449–484, 1988.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

C. A. Hosman, B. B. Hansen, and P. W. Holland. The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *The Annals of Applied Statistics*, 4(2):849–870, 2010.

J. Y. Hsu, D. S. Small, and P. R. Rosenbaum. Effect modification and design sensitivity in observational studies. *Journal of the American Statistical Association*, 108(501):135–148, 2013.

J. Y. Hsu, J. R. Zubizarreta, D. S. Small, and P. R. Rosenbaum. Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika*, 102(4):767–782, 2015.

P. J. Huber. *Robust Statistics*. Springer, New York, 1981.

IARC. *Some Traditional Herbal Medicines, Some Mycotoxins, Naphthalene and Styrene*, volume 82. World Health Organization, IARC Press, Lyon, France, 2002.

G. W. Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

G. W. Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

R. Jaeschke, G. Guyatt, H. Shannon, S. Walter, D. Cook, and N. Heddle. Basic statistics for clinicians: 3. assessing the effects of treatment: measures of association. *Canadian Medical Association Journal*, 152(3):351–357, 1995.

S. F. Jencks, M. V. Williams, and E. A. Coleman. Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428, 2009.

A. E. Jones. Point: Should Lactate Clearance Be Substituted for Central Venous Oxygen Saturation as Goals of Early Severe Sepsis and Septic Shock Therapy? Yes. *CHEST Journal*, 140(6):1406–1408, 2011.

T. K. Jones, B. D. Fuchs, D. S. Small, S. D. Halpern, A. Hanish, C. A. Umscheid, C. A. Baillie, M. P. Kerlin, D. F. Gaieski, and M. E. Mikkelsen. Post-acute care use and hospital readmission after sepsis. *Annals of the American Thoracic Society*, 12(6):904–913, 2015.

M. Jünger, T. M. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, G. Rinaldi, and L. A. Wolsey. *50 Years of Integer Programming 1958-2008*. Springer Science & Business Media, New York, 2009.

S. Karlin. *Mathematical Methods and Theory in Games, Programming, and Economics*, volume II. Dover, New York, 1992.

L. Keele, D. Small, and R. Grieve. Randomization based instrumental variables methods for binary outcomes with an application to the IMPROVE trial. available on lead author's website, 2014.

G. King and L. Zeng. The dangers of extreme counterfactuals. *Political Analysis*, 14(2): 131–159, 2006.

W. A. Knaus, E. A. Draper, D. P. Wagner, and J. E. Zimmerman. APACHE II: a severity of disease classification system. *Critical Care Medicine*, 13(10):818–829, 1985.

E. Lehmann. Nonparametric confidence intervals for a shift parameter. *The Annals of Mathematical Statistics*, 34(2):1507–1512, 1963.

E. Lehmann. *Elements of Large-Sample Theory*. Springer, New York, 2004.

M. Levy, J. Rapoport, S. Lemeshow, D. B. Chalfin, G. Phillips, and M. Danis. Association between critical care physician management and patient mortality in the intensive care unit. *Annals of Internal Medicine*, 148(11):801–809, 2008.

V. Liu, G. J. Escobar, J. D. Greene, J. Soule, A. Whippy, D. C. Angus, and T. J. Iwashyna. Hospital deaths in patients with sepsis from two independent cohorts. *Journal of the American Medical Association*, 312(1):90–92, 2014.

W. Liu, M. A. Brookhart, S. Schneeweiss, X. Mi, and S. Setoguchi. Implications of M bias in epidemiologic studies: a simulation study. *American Journal of Epidemiology*, 176(10): 938–948, 2012.

W. Liu, S. J. Kuramoto, and E. A. Stuart. An introduction to sensitivity analysis for

unobserved confounding in nonexperimental prevention research. *Prevention Science*, 14 (6):570–580, 2013.

N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

R. Marcus, P. Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

S. M. Marcus. Using omitted variable bias to assess uncertainty in the estimation of an aids education treatment effect. *Journal of Educational and Behavioral Statistics*, 22(2): 193–201, 1997.

F. Margot. Symmetry in integer linear programming. In *50 Years of Integer Programming 1958-2008*, pages 647–686. Springer, New York, 2010.

J. Maritz. A note on exact robust confidence intervals for location. *Biometrika*, 66(1): 163–170, 1979.

R. Mechanic. Post-acute care: the next frontier for controlling Medicare spending. *New England Journal of Medicine*, 370(8):692–694, 2014.

N. Meinshausen. Hierarchical testing of variable importance. *Biometrika*, 95(2):265–278, 2008.

M. E. Mikkelsen, A. N. Miltiades, D. F. Gaieski, M. Goyal, B. D. Fuchs, C. V. Shah, S. L. Bellamy, and J. D. Christie. Serum lactate is associated with mortality in severe sepsis independent of organ failure and shock. *Critical Care Medicine*, 37(5):1670–1677, 2009.

K. Ming and P. R. Rosenbaum. Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics*, 56(1):118–124, 2000.

D. Misselbrook and D. Armstrong. Patients' responses to risk information about the benefits of treating hypertension. *British Journal of General Practice*, 51(465):276–279, 2001.

J. E. Mitchell. Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of Applied Optimization*, pages 65–77, 2002.

H.-M. Nan, H. Kim, H.-S. Lim, J. K. Choi, T. Kawamoto, J.-W. Kang, C.-H. Lee, Y.-D. Kim, and E. H. Kwon. Effects of occupation, lifestyle and genetic polymorphisms of CYP1A1, CYP2E1, GSTM1 and GSTT1 on urinary 1-hydroxypyrene and 2-naphthol concentrations. *Carcinogenesis*, 22(5):787–793, 2001.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9 (in Polish). *Roczniki Nauk Roiniczych*, X:1–51, 1923. Reprinted in Statistical Science, 1990, 5(4):463-480.

Z. Obermeyer, M. Makar, S. Abujaber, F. Dominici, S. Block, and D. M. Cutler. Association between the medicare hospice benefit and health care utilization and costs for patients with poor-prognosis cancer. *Journal of the American Medical Association*, 312(18):1888–1896, 2014.

K. J. Ottenbacher, A. Karmarkar, J. E. Graham, Y.-F. Kuo, A. Deutsch, T. A. Reistetter, S. Al Snih, and C. V. Granger. Thirty-day hospital readmission following discharge from postacute rehabilitation in fee-for-service Medicare patients. *Journal of the American Medical Association*, 311(6):604–614, 2014.

C. Poole. On the origin of risk relativism. *Epidemiology*, 21(1):3–9, 2010.

R. Preuss, H. M. Koch, M. Wilhelm, M. Pischetsrieder, and J. Angerer. Pilot study on the naphthalene exposure of German adults and children by means of urinary 1-and 2-naphthol levels. *International Journal of Hygiene and Environmental Health*, 207(5):441–445, 2004.

M. A. Puskarich, S. Trzeciak, N. I. Shapiro, A. C. Heffner, J. A. Kline, and A. E. Jones. Outcomes of patients undergoing early sepsis resuscitation for cryptic shock compared with overt shock. *Resuscitation*, 82(10):1289–1293, 2011.

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

J. Rigdon and M. G. Hudgens. Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, (6):924 –935, 2014.

J. Rigdon and M. G. Hudgens. Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, 34(6):924 –935, 2015.

E. P. Rivers, R. Elkin, and C. M. Cannon. Counterpoint: should lactate clearance be substituted for central venous oxygen saturation as goals of early severe sepsis and septic shock therapy? No. *CHEST Journal*, 140(6):1408–1413, 2011.

J. M. Robins. Confidence intervals for causal parameters. *Statistics in Medicine*, 7(7): 773–785, 1988.

J. M. Rohde, A. J. Odden, C. Bonham, L. Kuhn, P. N. Malani, L. M. Chen, S. A. Flanders, and T. J. Iwashyna. The epidemiology of acute organ system dysfunction from severe sepsis outside of the intensive care unit. *Journal of Hospital Medicine*, 8(5):243–247, 2013.

J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005.

P. R. Rosenbaum. Dropping out of high school in the United States: An observational study. *Journal of Educational and Behavioral Statistics*, 11(3):207–224, 1986.

P. R. Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

P. R. Rosenbaum. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):597–610, 1991.

P. R. Rosenbaum. Detecting bias with confidence in observational studies. *Biometrika*, 79 (2):367–374, 1992.

P. R. Rosenbaum. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88(1):219–231, 2001.

P. R. Rosenbaum. *Observational Studies*. Springer, New York, 2002a.

P. R. Rosenbaum. Attributing effects to treatment in matched observational studies. *Journal of the American Statistical Association*, 97(457):183–192, 2002b.

P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17(3):286–327, 2002c.

P. R. Rosenbaum. Design sensitivity in observational studies. *Biometrika*, 91(1):153–164, 2004.

P. R. Rosenbaum. Heterogeneity and causality. *The American Statistician*, 59(1):147–152, 2005.

P. R. Rosenbaum. Sensitivity analysis for M-estimates, tests, and confidence intervals in matched observational studies. *Biometrics*, 63(2):456–464, 2007.

P. R. Rosenbaum. *Design of Observational Studies*. Springer, New York, 2010.

P. R. Rosenbaum. Optimal matching of an optimally chosen subset in observational studies. *Journal of Computational and Graphical Statistics*, 21(1):57–71, 2012.

P. R. Rosenbaum. Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics*, 69(1):118–127, 2013.

P. R. Rosenbaum. Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls. *Journal of the American Statistical Association*, 109(507):1145–1158, 2014.

P. R. Rosenbaum and A. M. Krieger. Sensitivity of two-sample permutation inferences in observational studies. *Journal of the American Statistical Association*, 85(410):493–498, 1990.

P. R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79(387): 516–524, 1984.

P. R. Rosenbaum and J. H. Silber. Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *Journal of the American Statistical Association*, 104(486):501–511, 2009.

D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

D. B. Rubin. Which ifs have causal answers. *Journal of the American Statistical Association*, 81(396):961–962, 1986.

D. B. Rubin. Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28(9):1420–1423, 2009.

D. B. Rubin and R. P. Waterman. Estimating the causal effects of marketing interventions using propensity score methodology. *Statistical Science*, 21(2):206–222, 2006.

J. J. Sabia. Does sex education affect adolescent sexual behaviors and health? *Journal of Policy Analysis and Management*, 25(4):783–802, 2006.

K. Sanctucci and B. Shah. Association of naphthalene with acute hemolytic anemia. *Academic Emergency Medicine*, 7(1):42–47, 2000.

T. Sato, S. Greenland, and J. M. Robins. On the variance estimator for the Mantel-Haenszel risk difference. *Biometrics*, 45(4):1323 – 1324, 1989.

E. Schechtman. Odds ratio, relative risk, absolute risk reduction, and the number needed to treat: which of these should we use? *Value in Health*, 5(5):431–436, 2002.

A. Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*. Springer, New York, 2003.

J. H. Silber, P. R. Rosenbaum, M. E. Trudeau, O. Even-Shoshan, W. Chen, X. Zhang, and R. E. Mosher. Multivariate matching and bias reduction in the surgical outcomes study. *Medical Care*, 39(10):1048–1064, 2001.

J. C. Sinclair and M. B. Bracken. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology*, 47(8):881–889, 1994.

E. A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010.

E. A. Stuart and K. M. Green. Using full matching to estimate causal effects in nonexperimental studies: examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, 44(2):395–406, 2008.

V. Sundararajan, C. M. MacIsaac, J. J. Presneill, J. F. Cade, and K. Visvanathan. Epidemiology of sepsis in Victoria, Australia. *Critical Care Medicine*, 33(1):71–80, 2005.

P. Suwan-ampai, A. Navas-Acien, P. T. Strickland, and J. Agnew. Involuntary tobacco smoke exposure and urinary levels of polycyclic aromatic hydrocarbons in the united states, 1999 to 2002. *Cancer Epidemiology Biomarkers & Prevention*, 18(3):884–893, 2009.

V. Todisco, J. Lamour, and L. Finberg. Hemolysis from exposure to naphthalene mothballs. *New England Journal of Medicine*, 325(23):1660–1661, 1991.

M. Traskin and D. S. Small. Defining the study population for an observational study to ensure sufficient overlap: a tree approach. *Statistics in Biosciences*, 3(1):94–118, 2011.

A. W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, Cambridge, 2000.

T. J. VanderWeele and O. A. Arah. Unmeasured confounding for general outcomes, treatments, and confounders: Bias formulas for sensitivity analysis. *Epidemiology*, 22(1):42–52, 2011.

T. J. VanderWeele and I. Shpitser. A new criterion for confounder selection. *Biometrics*, 67 (4):1406–1413, 2011.

T. J. VanderWeele and I. Shpitser. On the definition of a confounder. *Annals of Statistics*, 41(1):196–220, 2013.

N. Voigtländer and H.-J. Voth. Persecution perpetuated: the medieval origins of anti-semitic violence in nazi germany. *Quarterly Journal of Economics*, 127(3):1339–1392, 2012.

L. Wang and A. M. Krieger. Causal conclusions are most sensitive to unobserved binary covariates. *Statistics in Medicine*, 25(13):2257–2271, 2006.

M. Weitzman, S. Cook, P. Auinger, T. A. Florin, S. Daniels, M. Nguyen, and J. P. Winickoff. Tobacco smoke exposure is associated with the metabolic syndrome in adolescents. *Circulation*, 112(6):862–869, 2005.

P. H. Westfall and S. S. Young. *Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment*, volume 279. John Wiley & Sons, 1993.

S. A. Whittaker, B. D. Fuchs, D. F. Gaieski, J. D. Christie, M. Goyal, N. J. Meyer, C. Kean, D. S. Small, S. L. Bellamy, and M. E. Mikkelsen. Epidemiology and outcomes in patients with severe sepsis admitted to the hospital wards. *Journal of Critical Care*, 30(1):78–84, 2015.

F. Yang, J. R. Zubizarreta, D. S. Small, S. Lorch, and P. R. Rosenbaum. Dissonant conclusions when testing the validity of an instrumental variable. *The American Statistician*, 68 (4):253–263, 2014.

M. Yang, M. Koga, T. Katoh, and T. Kawamoto. A study for the proper application of urinary naphthols, new biomarkers for airborne polycyclic aromatic hydrocarbons. *Archives of Environmental Contamination and Toxicology*, 36(1):99–108, 1999.

B. Yu and J. L. Gastwirth. Sensitivity analysis for trend tests: application to the risk of radiation exposure. *Biostatistics*, 6(2):201–209, 2005.

J. R. Zubizarreta. 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, 2012.

J. R. Zubizarreta, M. Cerdá, and P. R. Rosenbaum. Effect of the 2010 Chilean earthquake on posttraumatic stress reducing sensitivity to unmeasured bias through study design. *Epidemiology*, 24(1):79–87, 2013.

J. R. Zubizarreta, R. D. Paredes, and P. R. Rosenbaum. 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, 2014.