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

To appear in the *Journal of the American Statistical Association: Theory and Methods*

Colin B. Fogarty Pixu Shi Mark E. Mikkelsen Dylan S. Small*

Abstract

We present methods for conducting hypothesis testing and sensitivity analyses for composite null hypotheses in matched observational studies when outcomes are binary. Causal estimands discussed include the causal risk difference, causal risk ratio, and the effect ratio. We show that inference under the assumption of no unmeasured confounding can be performed by solving an integer linear program, while inference allowing for unmeasured confounding of a given strength requires solving an integer quadratic program. Through simulation studies and data examples, we demonstrate that our formulation allows these problems to be solved in an expedient manner even for large data sets and for large strata. We further exhibit that through our formulation, one can assess the impact of various assumptions on the potential outcomes on the performed inference. `R` scripts are provided that implement our methodology.

Keywords: Causal Inference; Sensitivity Analysis; Integer Programming; Causal Risk; Effect Ratio

*Colin B. Fogarty is Doctoral Candidate, Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia PA 19104 (e-mail: `cfogarty@wharton.upenn.edu`). Pixu Shi is Doctoral Candidate, Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104. Mark E. Mikkelsen is Assistant Professor, Pulmonary, Allergy and Critical Care Division and Center for Clinical Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104. Dylan S. Small is Professor, Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia PA 19104.

1 Introduction

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, 2002b, 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, 2002a) 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.

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 A, 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, 2002b).

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, 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 (2002a) uses a similar approach for inference on the *attributable effect*, which is the effect of the treatment on the treated individuals. 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 worst-case p-value for this composite null. This is because the impact on the p-value by attributing a given outcome to the treatment 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. See Appendix B for further discussion on estimands which focus on the treatment effect on the treated versus the treatment effect on the entire study population.

Unfortunately, with other causal estimands of interest which do not solely focus on

the treatment's effect on treated individuals (the risk difference and risk ratio being two such estimands), and even for estimands focusing on the treated individuals but not assuming 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 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.

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.

2 Causal Inference after Matching

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 tha tthe 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, (Holland, 1988), 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. 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 covariates $\mathbf{x}_{ij}$ 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. 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 $\mathbb{P}(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathbf{Z} \in \Omega) = 1/|\Omega|$, where the notation $|B|$ denotes the number of elements in the set B.

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 (2002b, 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 (2002b, 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 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 (2002b, 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 (2002b, Section 4, Problem 12).

3 Composite Null Hypotheses

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 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 C 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.

3.2 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

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

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:

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

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 Fogarty et al. (2015) 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.

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 of Fogarty et al. (2015), which was catered towards maximizing the variance of the estimated causal risk difference under no unmeasured confounding. In Section 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 $(1, 1, ..., 1)$. 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 performed on the data in our motivating example from Section 1 returned 4893 strata, of which only 234 were unique.

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 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

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

and

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

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 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 $\boldsymbol{\psi}_{\theta_0} = [\psi(\theta_0)_{[11]1}, ..., \psi(\theta_0)_{[SP_S],\tilde{n}_S-1}]$.

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 3.2. 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 5.1), as well as a sensitivity analysis for any $\Gamma > 1$ (Section 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.

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 (P1) 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.

5.2 Sensitivity Analysis through Iterative Optimization

For $\Gamma > 1$, (P1) is instead an integer quadratic program. First, note that we reject the null with a two-sided alternative at size α if $(t_{\theta_0} - \boldsymbol{\mu}_{\theta_0}^T \mathbf{x})^2 / (\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}) \geq \chi_{1,1-\alpha}^2$ for all values of the potential outcomes that are consistent with the null being tested, where $\chi_{1,1-\alpha}^2$ is the $1-\alpha$ quantile of a χ_1^2 distribution. Equivalently, we need only determine whether $(t_{\theta_0} - \boldsymbol{\mu}_{\theta_0}^T \mathbf{x})^2 - \chi_{1,1-\alpha}^2 (\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}) \geq 0$ for all feasible $\mathbf{x}$. This can be done by minimizing (P1) with $\kappa = \chi_{1,1-\alpha}^2$ 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 (P1) at $\kappa = \chi_{1,1-\alpha}^2$ 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 (P1). To proceed, we find the value $\kappa = \kappa^*$ such that the minimal objective value of (P1) 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 (P1).
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 $\mathbf{x}_{\theta_0}^* = \mathbf{x}_{\theta_0}^{*(i)}$, 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 (P1) 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.

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 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 D, 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 4 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 (P1) by its continuous relaxation.

Appendix D 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 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

Table 1: 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.

Null Hypothesis; Confounder Strength	Avg. Time (s), Integer	Avg. Time (s), Relaxation	$\%(obj_{int} = obj_{rel})$	Avg Rel. 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%	$3 \times 10^{-4}\%$
$\varphi = 1.75; \Gamma = 1$	5.86	5.62	0%	$4 \times 10^{-5}\%$
$\varphi = 1.75; \Gamma = 3$	10.85	7.82	3.2%	0.002%

an equivalent, but highly symmetric, formulation in order to highlight the importance of avoiding symmetry for achieving a strong formulation with reasonable computation time. Finally, we 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.

6 Data Examples

We employ our methodology in two data examples. In Section 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 1, wherein we compare hospital readmission rates for two different post-hospitalization protocols after an acute care hospitalization. In Section 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 $\hat{\theta} := \textbf{SOLVE}\{\theta : T(\theta) = 0\}$; see Appendix E 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 (2002b) 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.

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 2. 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 design sensitivity, but also computation time. The computations for each value of $\Gamma > 1$ shown took an average of 1.5 seconds under the assumption of nonnegativity, but 75 seconds without this assumption. See Appendix F 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.

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

Table 2: Sensitivity analysis for an 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

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 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 3, 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 F 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

Table 3: 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

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.

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 5.3 and Appendix D 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 G 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.

APPENDIX

A Balance on Observed Covariates in Our Motivating Example

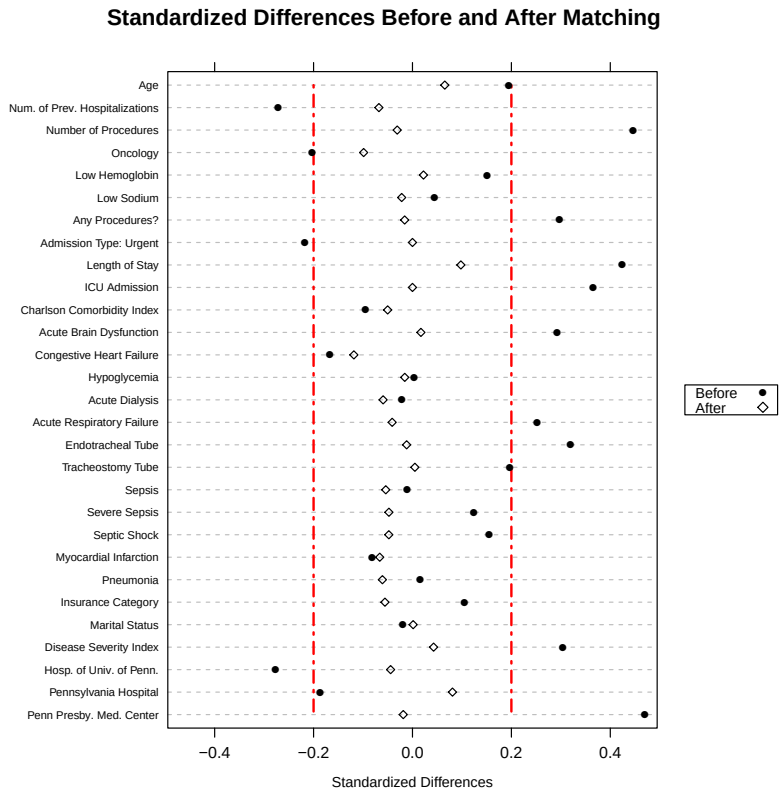

Figure 1: Covariate Imbalances Before and After Matching. The dotplot (a Love plot) shows the absolute standardized differences without matching, and after conducting a matching with a variable number of controls. The vertical dotted line corresponds to a standardized difference threshold of 0.2, which is often regarded as the maximal allowable absoute 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 Treatment Effects and Treatment Effects on the Treated

In determining whether interest should lie in the attributable effect of Rosenbaum (2001, 2002a) rather than the risk difference or risk ratio, one must decide whether the question of interest concerns the treatment effect on the entire study population, or rather only

concerns the effect of the intervention for those individuals who were treated. The attributable effect is itself a treatment effect *on the treated*, in that it is defined as $A = \sum_{i=1}^{I} \sum_{j=1}^{n_i} Z_{ij}(r_{Tij} - r_{Cij})$, whereas the risk difference and risk ratio concern the effects of the treatment for the entire study population. While the group of treated individuals will look the same as the overall study population in a randomized clinical trial, these two groups of individuals can be quite different in observational studies due to self-selection. Thus, these estimands speak to different sets of individuals in the study population.

In the example presented in Section 1.2 on hospital readmission rates, if we define the treatment group as the individuals who went to acute care rehabilitation and assign a response of 1 (a "success") to individuals did not require hospital readmission, the attributable effect then compares the number of individuals *assigned an acute rehabilitation center* who were not rehospitalized to what that number would have been if all the individuals who went to acute rehabilitation actually used home with home health services for their rehabilitation. If we additionally assume a nonnegative treatment effect, the attributable effect would have an additional interpretation as the number of individuals assigned to an acute rehabilitation center who avoided rehospitalization because of the acute rehabilitation center (i.e., who would have been rehospitalized had they instead utilized home with home health services for their rehabilitation). The risk difference, on the other hand, would compare the proportion of individuals in the *entire study population* who would have avoided rehospitalization if all individuals had been assigned to acute rehabilitation to the proportion avoiding rehospitalization if all individuals utilized home with home health services.

The determination of which target population (the entire study population versus only the individuals who received the treatment) is most relevant often depends on the research question being investigated. Austin (2010) notes that "applied researchers should decide whether the [average treatment effect] or the [average treatment effect on the treated] is of greater utility or interest in their particular research context." He then provides two examples to illustrate why the estimand of greatest relevance dependence is not uniform across applications. He notes that if one where investigating the effectiveness of an intensive and structured smoking cessation program, the treated individuals alone may be of greater interest as many smokers may not be interested in the program due to its intensive nature. On the other hand, if the program instead involved physicians giving brochures to patients, then the effect on the *entire* population of smokers (both those receiving and those not receiving the brochure) may be of greater interest as there are minimal barriers to a patient receiving the treatment (the brochure). Heckman and Robb (1985) and Heckman et al. (1997) further argue that the treated units themselves are often of more interest than the overall population if the intervention is narrowly targeted (e.g., if it is such that most controls would never consider undergoing the intervention).

Investigation of the average treatment effect (risk difference) has long been a pursuit of great interest in randomized experiments (Neyman, 1923), and this interest has carried over to the analysis of observational studies. For example, Imbens (2004) notes that the average treatment effect is the most commonly studied causal estimand in econometric literature. Investigating the attributable effect while also assuming a known direction of effect can remove the need for integer programming in conducting inference and sensitivity analyses (Rosenbaum, 2002a); however, one must be sure that

the attributable effect is the most appropriate estimand for the problem at hand before pursuing inference on it. This determination should not be made solely on the basis of computational complexity. In many cases the risk difference and risk ratio may be more appropriate, and in these instances the methodology presented in this work can facilitate inference and sensitivity analyses.

C 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) take a pragmatic approach and suggest 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), Sinclair and Bracken (1994) for further discussion of this matter.

D 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 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 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 matched 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.

D.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 2 demonstrates, the time for the optimization routine itself appears to increase with I, the number of matched sets. Figure 2 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 3 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 $\sum_{s=1}^{S}(\tilde{n}_s - 1)\prod_{(z,r,d)\in\{0,1\}^3}(|\mathcal{T}_s^{zrd}|+1)^2 = 2\times(4\times 3^2\times 2^2+32\times 2^6)$. 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.

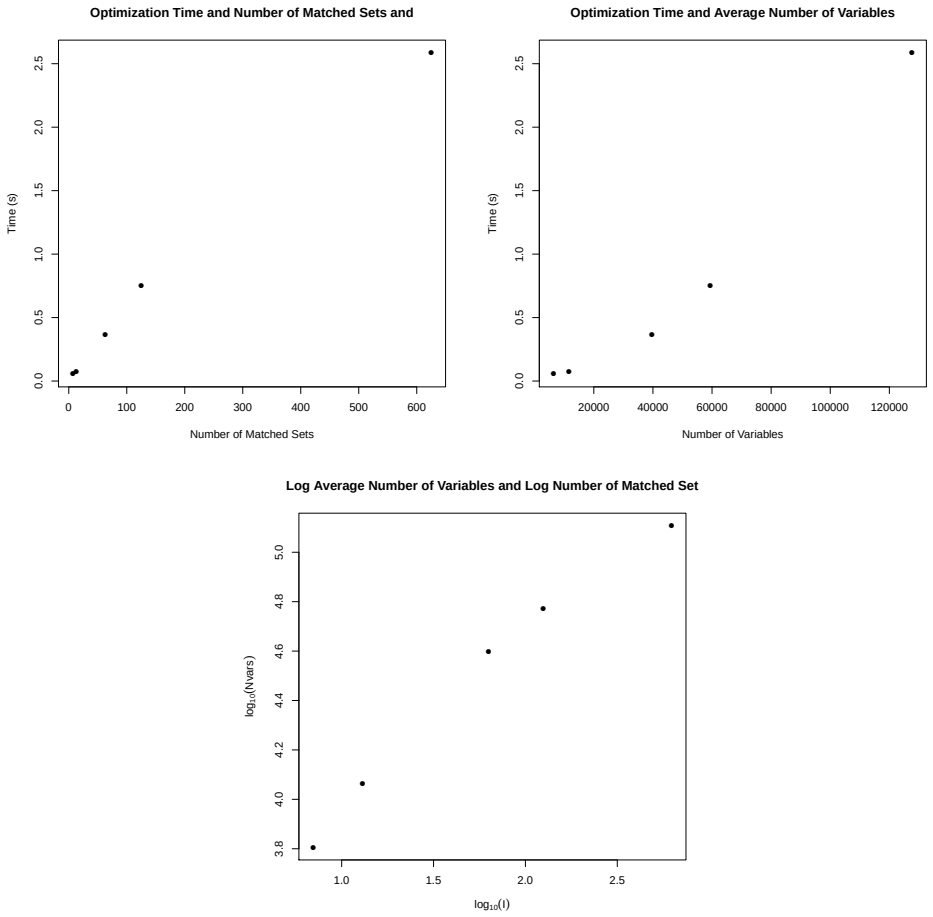

Figure 2: (Top-left) Optimization time and the number of matched sets; (top-right) optimization time and the number of optimization variables; and (bottom) log number of matched sets and log-number of optimization variables.

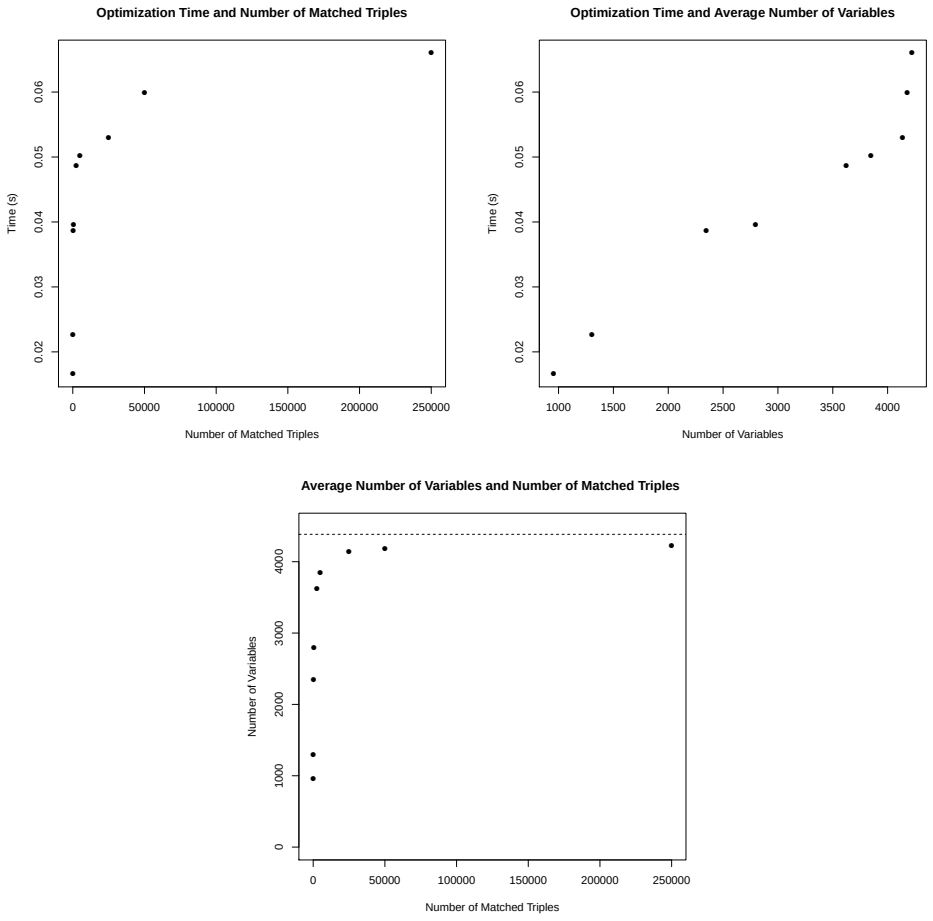

Figure 3: (Top-left) Optimization time and the number of matched triples; (top-right) optimization time and the average number of variables; and (bottom) average number of variables and the number of matched triples.

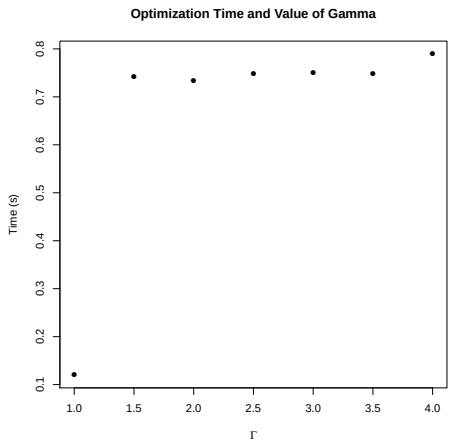

Figure 4: Optimization time and value of Γ.

D.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 4 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.

D.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 5, 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.

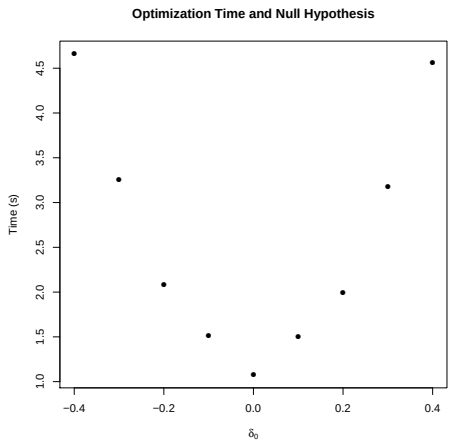

Figure 5: Optimization time and null hypothesis being tested. The true risk difference was set to zero throughout

D.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 6, we see that 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 6 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.

D.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 7, 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 8, 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 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

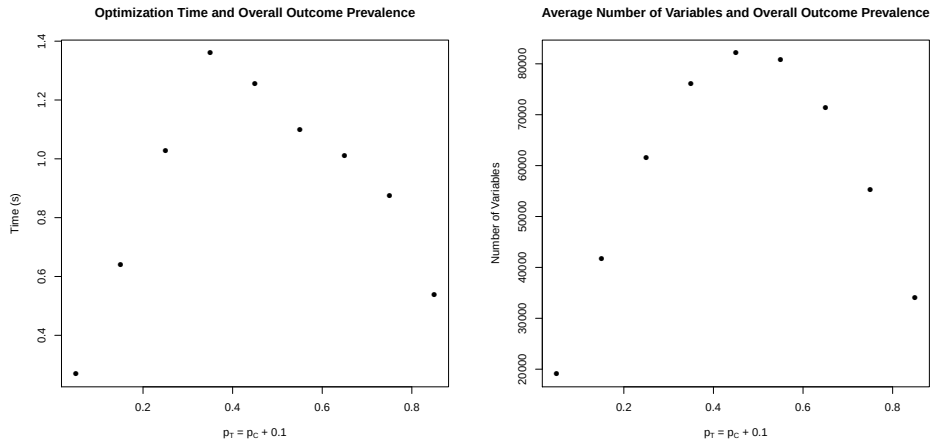

Figure 6: (Left) Optimization time and overall outcome prevalence; and (right) number of variables and outcome prevalence.

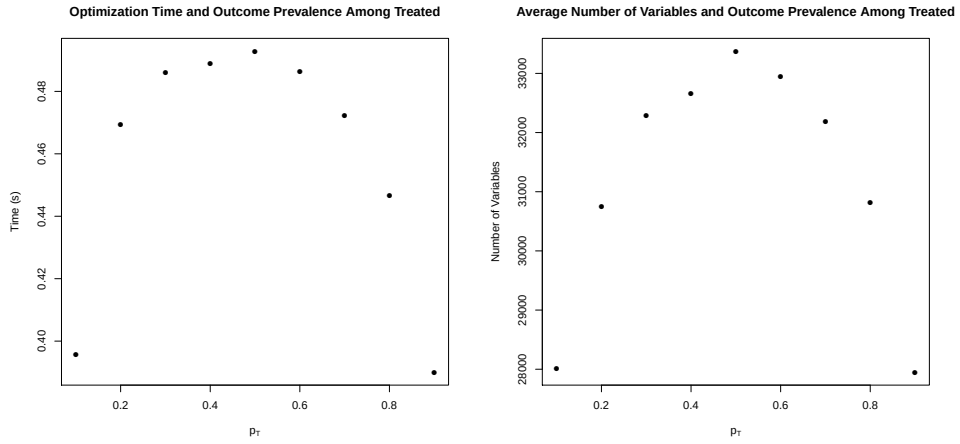

Figure 7: (Left) Optimization time and outcome prevalence under treatment; and (right) number of variables and outcome prevalence under treatment.

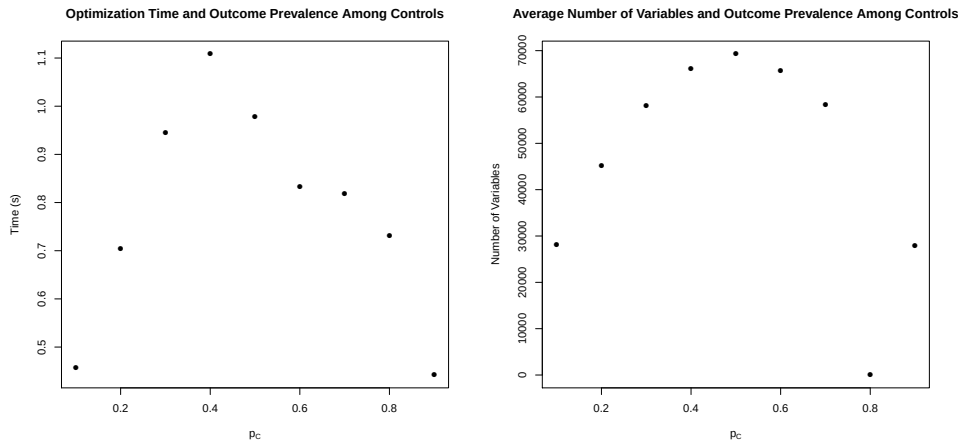

Figure 8: (Left) Optimization time and outcome prevalence under control; and (right) number of variables and outcome prevalence under control.

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.

D.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 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 Fogarty et al. (2015). This full match yielded strata of maximal size 8, representing a substantially easier optimization problem than the one presented in Section 5.3 of the manuscript. 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.

D.7 Simulation Using Actual Data

In each of 1000 iterations we sample 1250 matched sets from the strata in our motivating example from Section 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 4 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 4: 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; Confounder Strength	Avg. Time (s), Integer	Avg. Time (s), Relaxation	$\%(obj_{int} = obj_{rel})$	Avg Rel. Gap If Different
$\delta = 0.05; \Gamma = 1.00$	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%

E 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, $\hat{\theta} := \textbf{SOLVE}\{\theta : T(\theta) = 0\}$. 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 Fogarty et al. (2015) for a discussion of the latter point as it pertains to constructing confidence intervals for the risk difference within a matched observational study.

F Assuming a Known Direction of Effect Impacts Reported Sensitivity

In both examples in Section 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 strata 1, suppose $R_{11} = R_{12} = 1$, while in strata 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.

G 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 strata j. In the notation of Section 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 4.1 in the manuscript, 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 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 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.

G.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 strata 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.

References

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

Austin, P. C. (2010). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in Medicine*, 29(20):2137–2148.

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

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

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

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

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

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

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

Fogarty, C. B., Mikkelsen, M. E., Gaieski, D. F., and Small, D. S. (2015). Discrete optimization for interpretable study populations and randomization inference in an observational study of severe sepsis mortality. *Journal of the American Statistical Association*, to appear.

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

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

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

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

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

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

Heckman, J. and Robb, R. (1985). Using longitudinal data to estimate age, period and cohort effects in earnings equations. In *Cohort Analysis in Social Research*, pages 137–150. Springer.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Ottenbacher, K. J., Karmarkar, A., Graham, J. E., Kuo, Y.-F., Deutsch, A., Reistetter, T. A., Al Snih, S., and Granger, C. V. (2014). 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.

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

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

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

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

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

Rosenbaum, P. R. (2002b). *Observational Studies*. Springer, New York.

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

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

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

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

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

Schechtman, E. (2002). 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.

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

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

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

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