Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

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

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

To cite this article: Colin B. Fogarty, Pixu Shi, Mark E. Mikkelsen & Dylan S. Small (2016): Randomization Inference and Sensitivity Analysis for Composite Null Hypotheses with Binary Outcomes in Matched Observational Studies, Journal of the American Statistical Association, DOI: 10.1080/01621459.2016.1138865

To link to this article: http://dx.doi.org/10.1080/01621459.2016.1138865

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

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 about the potential outcomes on the performed inference. `R` scripts are provided that implement our methods.

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

combinations of potential outcomes, even *without* considering values for the unmeasured confounder.

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

Unfortunately, in the absence of a known direction of effect finding the worst-case allocation does not simplify in the same manner. This is because finding the potential outcome allocation with the largest impact on the *p*-value on a stratum-wise basis does not readily yield an allocation that satisfies the composite null. The problem is not separable on a stratum-wise basic even asymptotically, as the requirement that the composite null must be true necessarily links the strata together in a complex manner. There are two non-complementary forces at play in the required optimization problem: for some strata, the potential outcome allocations should maximize the impact on the *p*-value, while in other strata the missing potential outcome allocations should work towards satisfying the composite null. For our motivating example, there are over 300,000 types of contributions to the *p*-value that must be considered in the

sensitivity analysis when we do not assume a known direction of effect (as is shown in Section 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 that the treatments for all individuals with the same level of observed treatment are truly comparable (Rubin, 1986). Let Z_{ij} be an indicator variable that takes the value 1 if individual j in stratum i is assigned to the treatment. Each individual has two sets of binary potential outcomes: one under treatment, $\{r_{Tij}, d_{Tij}\}$, and one under control, $\{r_{Cij}, d_{Cij}\}$. r_{Tij} and r_{Cij} are the primary outcomes of interest, while d_{Tij} and d_{Cij} are indicators of whether or not an individual would actually take the treatment when randomly assigned to the treatment or control group. The observations for each individual are $R_{ij} = r_{Tij}Z_{ij} + r_{Cij}(1 - Z_{ij})$ and $D_{ij} = d_{Tij}Z_{ij} + d_{Cij}(1 - Z_{ij})$; see Neyman (1923) and Rubin (1974) for more on the potential outcomes framework. In the classical experimental setting, $d_{Tij} - d_{Cij} = 1\ \forall i, j$, and hence all individuals take the administered treatment. For a randomized encouragement design, Z_{ij} represents the encouragement to take the treatment (which is randomly assigned to patients), while d_{Tij} and d_{Cij} are the actual treatment received if $Z_{ij} = 1$ and $Z_{ij} = 0$ respectively (Holland, 1988). Matched observational studies assuming strong ignorability (Rosenbaum and Rubin, 1983) aim to replicate a classical stratified experiment, whereas matched studies employing an instrumental variable strive towards a randomized encouragement design, with Z_{ij} being the instrumental variable.

There are $N = \sum_{i=1}^{I} n_i$ total individuals in the study. Each individual has observed 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, where the notation $|B|$ denotes the number of elements in the set B. Hence, quantities dependent on the assignment vector such as $\mathbf{Z}$, $\mathbf{R}$ and $\mathbf{D}$ are random, whereas $\mathcal{F} = \{\mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C, \mathbf{X}, \mathbf{u}\}$

contains fixed quantities. For a randomized experiment, $\mathbb{P}(Z_{ij} = 1|\mathcal{F}, \mathbf{Z} \in \Omega) = m_i/n_i$, and $\mathbb{P}(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathbf{Z} \in \Omega) = 1/|\Omega|$.

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 $\log\left(\pi_{ij}/(1-\pi_{ij})\right) = \zeta(\mathbf{x}_{ij}) + \gamma u_{ij}$, 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 prob-

ability, $\mathbb{P}(T \geq t)$, 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 B 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

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

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:

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

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

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

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

Under mild regularity conditions, Lyapunov's central limit theorem yields that all three of the test statistics $T(\theta_0)$ under consideration are well approximated by a normal distribution for $\Gamma \geq 1$. See 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 $(1, 1, ..., 1)$. An analogous argument shows that there are only $|\mathcal{T}_i^{000}| + 1$ non-exchangeable arrangements for d_{Tij}, thus resulting in $(|\mathcal{T}_i^{000}| + 1)^2$ total non-exchangeable allocations. The same logic yields a contribution of $(|\mathcal{T}_i^{zrd}| + 1)^2$ for each of the other seven partitions.

Additional structure is often imposed on the potential outcomes on top of consistency with

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

It would seem as though we must consider at most $\prod_{i=1}^{I}\prod_{(z,r,d)\in\{0,1\}^3}(|\mathcal{T}_i^{zrd}| + 1)^2$ different distributions for $T(\theta_0) = \sum_{i=1}^{I} T_i(\theta_0)$ in our optimization problem. Fortunately, note first that we assume independence between strata, and further note that we are using a normal approximation to conduct inference. Hence, both the expectations and variances *sum* between strata and we do not need to consider covariances between strata. Further, in the same way that there were a limited number of non-exchangeable allocations of potential outcomes in each stratum due to repetition, many observed 2^3 factorial tables in the data are repeated multiple times. For example, the matching with multiple controls described in Section 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 $C_i = (|\mathcal{T}_i^{000}|, ..., |\mathcal{T}_i^{111}|)$ be the observed counts of the 2^3 tables for stratum i. $\mathfrak{C} = \{C_1, ..., C_I\}$ 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 $C_i = C_\ell$. 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 $\boldsymbol{\mu}_{\theta_0} = [\mu(\theta_0)_{[11]1}, ..., \mu(\theta_0)_{[S P_S],\tilde{n}_S - 1}]$, and let $\boldsymbol{\nu}_{\theta_0} = [\nu(\theta_0)_{[11]1}, ..., \nu(\theta_0)_{[S P_S],\tilde{n}_S - 1}]$. 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 $\boldsymbol{\psi}_{\theta_0} = [\psi(\theta_0)_{[11]1}, ..., \psi(\theta_0)_{[S P_S],\tilde{n}_S - 1}]$.

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 $\mathbf{x} = [x_{[11]1}, .., x_{[S\,P_s],\tilde{n}_S-1}]$. 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):

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

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 $z_{\theta_0} = t_{\theta_0} / \sqrt{\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}_{\theta_0}^*}$.

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 $\theta_\ell = \textbf{SOLVE}\left\{\theta : t_\theta / \sqrt{\boldsymbol{\nu}_\theta^T \mathbf{x}_\theta^*} = z_{1-\alpha/2}\right\}$ and $\theta_u = \textbf{SOLVE}\left\{\theta : t_\theta / \sqrt{\boldsymbol{\nu}_\theta^T \mathbf{x}_\theta^*} = z_{\alpha/2}\right\}$, 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 $(t_{\theta_0} - \boldsymbol{\mu}_{\theta_0}^T \mathbf{x})^2 / (\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}) \geq \chi^2_{1,1-\alpha}$ for all values of the potential outcomes that are consistent with the null being tested, where $\chi^2_{1,1-\alpha}$ is the $1-\alpha$ quantile of a χ^2_1 distribution. Equivalently, we need only determine whether $(t_{\theta_0} - \boldsymbol{\mu}_{\theta_0}^T \mathbf{x})^2 - \chi^2_{1,1-\alpha}(\boldsymbol{\nu}_{\theta_0}^T \mathbf{x}) \geq 0$ for all feasible $\mathbf{x}$. This can be done by minimizing (P1) with $\kappa = \chi^2_{1,1-\alpha}$ over all feasible $\mathbf{x}$, and checking whether or not the objective value at $\mathbf{x}_{\theta_0}^*$ is greater than zero.

One may also be interested in knowing the worst-case deviate itself (equivalently, the worst-case p-value), rather than simply knowing the result of the test. The optimal vector $\mathbf{x}_{\theta_0}^*$ for (P1) at $\kappa = \chi^2_{1,1-\alpha}$ need not result in the worst-case deviate; however, we now show that

we can find the worst-case p-value through an iterative procedure based on (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 $\kappa^{(i)} = (t_{\theta_0} - (\boldsymbol{\mu}^T_{\theta_0}\mathbf{x}^{*(i)}_{\theta_0}))^2/(\boldsymbol{\nu}^T_{\theta_0}\mathbf{x}^{*(i)}_{\theta_0})$.
4. If $\kappa^{(i)} = \kappa^{(i-1)}$ terminate the algorithm: set $\mathbf{x}^*_{\theta_0} = \mathbf{x}^{*(i)}_{\theta_0}$, and set $\kappa^* = \kappa^{(i)}$.
5. Otherwise, return to step 2. Repeat until convergence.

Note that the sequence $\{\kappa^{(i)}\}$ is bounded below by 0. It is also monotone decreasing for $i \geq 1$, as $(t_{\theta_0} - \boldsymbol{\mu}^T_{\theta_0}\mathbf{x}^{*(i+1)}_{\theta_0})^2 - \kappa^{(i)}(\boldsymbol{\nu}^T_{\theta_0}\mathbf{x}^{*(i+1)}_{\theta_0}) \leq (t_{\theta_0} - \boldsymbol{\mu}^T_{\theta_0}\mathbf{x}^{*(i)}_{\theta_0})^2 - \kappa^{(i)}(\boldsymbol{\nu}^T_{\theta_0}\mathbf{x}^{*(i)}_{\theta_0}) = 0$, which implies $\kappa^{(i)} \geq (t_{\theta_0} - (\boldsymbol{\mu}^T_{\theta_0}\mathbf{x}^{*(i+1)}))^2/(\boldsymbol{\nu}^T_{\theta_0}\mathbf{x}^{*(i+1)}_{\theta_0}) = \kappa^{(i+1)}$. 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 C, 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 1 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 sim-

ulations 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 C 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 an equivalent, but highly symmetric, formulation in order to highlight the importance of avoiding symmetry for achieving a strong formulation with reasonable computation time. We also present a simulation study akin to the one presented in this section but using real data for the outcome variables as opposed to simulated outcomes. Finally, we provide advice for using our procedure under time constraints for the optimization routine.

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 D 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 as-

sumption 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 robustness against unmeasured confounding, but also computation time. The computations for each value of $\Gamma > 1$ shown took an average of 1.5 seconds under the assumption of non-negativity, but 75 seconds without this assumption. See Appendix E 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 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 E for further discussion. Because of this, the assumptions of monotonicity and the exclusion restriction cannot impact the sensitivity analysis at $\lambda = 0$ unless non-negativity is not enforced. Furthermore, without assuming a direction of effect, monotonicity can only affect the performed inference if it is enforced in concert with the exclusion restriction at $\lambda = 0$ and vice versa. For $\lambda = 0.1$, the test no longer corresponds exclusively to one of Fisher's sharp null when non-negativity is imposed. We thus see that each assumption impacts the study's robustness against unmeasured confounding to varying degrees. For all combinations of assumptions and each value of Γ tested, the corresponding integer quadratic program solved in under 2 seconds.

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 C 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 F 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.

SUPPLEMENTARY MATERIAL

Appendices Appendix A shows the standardized differences before and after matching in our example in Section 1.2. Appendix B compares and contrasts the risk difference and the risk ratio. Appendix C contains simulation studies on the computation time of our integer program. Appendix D discusses point estimation through the use of *m*-estimators for our causal estimands. Appendix E elucidates the impact that assuming a know direction of effect can have on the performed inference (.pdf file). Appendix F discusses how our methodology can be used to perform sensitivity analyses for simple null hypotheses with general outcomes.

R-scripts `compositeBinary.R` provides code for hypothesis testing, confidence intervals, and sensitivity analysis for the causal risk difference, causal risk ratio, and effect ratio

in matched observational studies under a host of assumptions on the potential outcomes. `sensitivitySimple.R` provides code for performing sensitivity analyses in matched observational studies for simple null hypotheses with test statistics of the form $\mathbf{Z}^T\mathbf{q}$ (.R file).

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.

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.

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.

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.

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.

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.

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.

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.

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. (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 nonrandomized 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.

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.

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%	3×10^{-4}%
$\varphi = 1.75; \Gamma = 1$	5.86	5.62	0%	4×10^{-5}%
$\varphi = 1.75; \Gamma = 3$	10.85	7.82	3.2%	0.002%

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	$r_{Tij} \gtrless r_{Cij}$	$r_{Tij} \leq r_{Cij}$
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

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