# MULTIVARIATE ONE-SIDED TESTING IN MATCHED OBSERVATIONAL STUDIES AS AN ADVERSARIAL GAME

**Peter L. Cohen**
Operations Research Center
Massachusetts Institute of Technology
1 Amherst Street
Cambridge Massachusetts 02142, U.S.A
plcohen@mit.edu

**Matt A. Olson**
The Voleon Group
2170 Dwight Way
Berkeley, CA 94704, U.S.A
molson@voleon.com

**Colin B. Fogarty**
Operations Research and Statistics Group
Massachusetts Institute of Technology
100 Main Street
Cambridge, Massachusetts 02142, U.S.A
cfogarty@mit.edu

December 20, 2018

## ABSTRACT

We present a multivariate one-sided sensitivity analysis for matched observational studies, appropriate when the researcher has specified that a given causal mechanism should manifest itself in effects on multiple outcome variables in a known direction. The test statistic can be thought of as the solution to an adversarial game, where the researcher determines the best linear combination of test statistics to combat nature's presentation of the worst-case pattern of hidden bias. The corresponding optimization problem is convex, and can be solved efficiently even for reasonably sized observational studies. Asymptotically the test statistic converges to a chi-bar-squared distribution under the null, a common distribution in order restricted statistical inference. The test attains the largest possible design sensitivity over a class of coherent test statistics, and facilitates one-sided sensitivity analyses for individual outcome variables while maintaining familywise error control through is incorporation into closed testing procedures.

## 1 On Multiplicity and Causality

Controlled randomization protects empirical evidence against a host of counterclaims. A significant finding may well be due to random chance alone, but cannot be dismissed on the grounds of biases unaccounted for by the study's design. Observational evidence provides no such assurance, and causal inference in observational studies involves ambiguity which randomization eschews: Is the association an effect, or is it bias from self-selection? Anticipating skepticism, a practitioner may take measures when planning an observational study to properly frame the debate, rendering certain criticism unwarranted should the practitioner's hypothesis be true. While ambiguity cannot be eliminated, quasi-experimental devices may be employed to help clarify the step from association to causation in observational studies; see Shadish et al. [2002] and Rosenbaum [2015] for an overview. One such device, known alternatively as pattern specificity, multiple operationalism, or coherence, advocates that observational studies be designed with the objective of confirming a complex pattern of predictions made by the causal theory in question. This is in keeping with Fisher's notion of elaborate theories, which advocates that the practitioner "envisage as many different consequences of

[a causal hypothesis's] truth as possible, and plan observational studies to discover whether each of these consequences is found to hold" [Cochran, 1965, §5, p. 252]. Complex predictions imperil the practitioner's hypothesis, as doubt is cast should any prediction fail in the observational study at hand. Should the evidence prove coherent with the theory's predictions, fortification is provided as attributing a complex pattern to hidden bias requires that hidden bias could reproduce the particular pattern of association.

One way in which a theory can be made elaborate is through predicting that an intervention will affect multiple outcome variables in a prespecified direction. While the practitioner hopes that each prediction holds, should certain predictions fail she would regardless like to quantify which components came to fruition as a means of refining understanding of the mechanism in question. With this comes the attending issues of multiple comparisons. Concerns over a loss in power from multiplicity control may lead practitioners to instead investigate the outcome they believe *a priori* will be most affected, reducing the extent to which Fisher's advice is followed.

The qualitative benefits of multiple outcomes in observational studies are thus at odds with the statistical corrections they require. This tension exists not only when assuming no hidden bias, but also in the sensitivity analysis where the researcher quantifies the magnitude of hidden bias required to overturn the study's conclusions. In what follows, we present a new method for sensitivity analysis in multivariate one-sided testing, appropriate when the researcher anticipates a particular direction of effects for multiple outcome variables. The test adaptively combines outcome-specific test statistics, has the optimal design sensitivity over a class of multivariate tests respecting coherence, and leads to substantial improvements in power when the researcher's prediction proves correct. The method greatly attenuates the impact of multiplicity control on power for testing individual outcome variables through its use in closed testing procedures [Marcus et al., 1976], facilitating the analysis of multiple outcomes for demonstrating coherence.

## 2 Hidden Bias in Matched Observational Studies

### 2.1 A finely stratified experiment with multiple outcomes

There are $I$ independent strata, the $i$th of which contains $n_i \geq 2$ individuals. Individual $j$ in stratum $i$ has a $P$-dimensional vector of observed covariates $x_{ij}$, along with an unobserved covariate $u_{ij}$, $0 \leq u_{ij} \leq 1$. The strata are formed such that $x_{ij} \approx x_{ij'}$ for any two individuals $j \neq j'$ in stratum $i$; however, $u_{ij}$ are unobserved, it is likely that $u_{ij} \neq u_{ij'}$. We take $Z_{ij}$ as the indicator of treatment for the $j$th individual in stratum $i$, such that $Z_{ij} = 1$ if assigned to treatment and $Z_{ij} = 0$ otherwise. Each strata contains one treated individual and $n_i - 1$ controls such that $\sum_{j=1}^{n_i} Z_{ij} = 1$ $(i = 1, ..., I)$. See Fogarty [2018] for more on this particular class of stratified experiments, therein referred to as finely stratified experiments.

Each individual has two vectors of potential outcomes of length $K$: the responses for each outcome variable under control $r_{Cij} = (r_{Cij1}, ..., r_{CijK})^T$, and the responses under treatment $r_{Tij} = (r_{Tij1}, ..., r_{TijK})^T$. The $K$-dimensional vector of treatment effects $\tau_{ij} = r_{Tij} - r_{Cij}$ is not observed; instead, we observe the vector $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$. Let $Z = (Z_{11}, ..., Z_{In_I})^T$ be the lexicographically ordered vector of treatment assignments of length $N$, and let the analogous hold for $u$ along with $r_{Ck}, r_{Tk}$ and $R_k$ for $k = 1, ..., K$. The $N \times K$ matrix with lexicographically ordered rows containing $R_{ij}^T$ is $R$.

Let $\mathcal{F} = \{r_{Cij}, r_{Tij}, x_{ij}, u_{ij} : i = 1, ..., I; j = 1, ..., n_i\}$ be a set containing the potential outcomes along with the measured and unmeasured covariates for each individual in the observational study. In what follows we consider inference conditional upon $\mathcal{F}$, such that a generative model for the potential outcomes is neither assumed nor required. Let $\Omega = \{z : \sum_{j=1}^{n_i} z_{ij} = 1; i = 1, ..., I\}$ be the set of $\prod_{i=1}^{I} n_i$ treatment assignments adhering to the stratified design, and let $\mathcal{Z} = \{Z \in \Omega\}$ be the event that the observed treatment assignment satisfies this design. In a finely stratified experiment $\mathrm{pr}(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z}) = 1/n_i$ and $\mathrm{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ where $|A|$ is the cardinality of the set $A$.

### 2.2 A model for biased treatment assignment

Matched observational studies aim to mimic the finely stratified experiment described in §2.1. Matching algorithms assign individuals to matched sets on the basis of observed covariates such that $x_{ij} \approx x_{ij'}$ for individuals $j$ and $j'$ in the same matched set $i$; see Hansen [2004] and Zubizarreta [2012] among many for more on matching algorithms and the optimization problems underpinning them. A simple model for treatment assignment in an observational study states that before matching, individuals are assigned to treatment independently with unknown probabilities $\pi_{ij} = \mathrm{pr}(Z_{ij} = 1 \mid \mathcal{F})$. While one may hope that $\pi_{ij} \approx \pi_{ij'}$ after matching, proceeding as such would be specious due to the potential presence of unmeasured confounding. The model of Rosenbaum [2002, Chapter 4] stipulates that

individuals in the same matched set may differ in their odds of assignment to treatment by at most a factor of $\Gamma$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1 - \pi_{ij'})}{\pi_{ij'}(1 - \pi_{ij})} \leq \Gamma. \tag{1}$$

The parameter $\Gamma$ controls the degree to which unmeasured confounding may have impacted the treatment assignment process. The value $\Gamma = 1$ returns a randomized finely stratified experiment, while $\Gamma > 1$ allows the unobserved covariates to tilt the randomization distribution to a degree controlled by $\Gamma$. For instance, $\Gamma = 2$ stipulates that individuals in the same matched set might truly differ in their odds of receiving the treatment by a factor of at most two. Returning attention to the matched structure by conditioning on $\mathcal{Z}$, this model is equivalent to assuming

$$\mathrm{pr}\left(Z = z \mid \mathcal{F}, \, \mathcal{Z}\right) = \frac{\exp\left(\gamma z^T u\right)}{\sum_{b \in \Omega} \exp\left(\gamma b^T u\right)}, \tag{2}$$

where $\gamma = \log(\Gamma)$ and $u$ lies in the $N$-dimensional unit cube, call it $\mathcal{U}$; see Rosenbaum [1995] or Rosenbaum [2002, Chapter 4] for a proof of this equivalence.

### 2.3 Sensitivity analysis for a particular outcome

Assume without loss of generality that the outcomes have been recorded such that positive values for the treatment effects $\tau_{ijk}$ are predicted by the causal theory under study. For each outcome variable, we consider tests of the null hypothesis of non-positive treatment effects,

$$H_k : r_{Tijk} \leq r_{Cijk} \quad (i = 1, ..., I; j = 1, ..., n_i).$$

$H_k$ is a composite null hypothesis. Elements of $H_k$ include the null of a non-positive constant effect for all individuals, $r_{Tijk} = r_{Cijk} + \delta_k$ for any scalar $\delta_k \leq 0$. Fisher's sharp null of no effect is $\delta_k = 0$, thus representing the boundary of $H_k$.

We consider test statistics for each outcome variable which are effect increasing sum statistics. Sum statistics are statistics of the form $T_k(Z, R_k) = Z^T q_k$ where $q_k = q_k(R_k)$ is a pre-specified function of the observed responses $R_k$. A test statistic is effect increasing if $T_k(z, r_k^*) \geq T_k(z, r_k)$ whenever $(2z_{ij} - 1)(r_{ijk}^* - r_{ijk}) \geq 0$ for all $i$ and $j$. In words, this means that if every treated unit did better with $r_k^*$ than with $r_k$, and if every control did worse with $r_k^*$ than with $r_k$, then the test statistic corresponding to the observed outcomes $r_k^*$ would be larger than it would have been under $r_k$. Most familiar test statistics, including differences-in-means, rank tests, and $m$-tests are endowed with these properties; see Rosenbaum [2002, Chapter 2.4.4] and Rosenbaum [2016, §3.1] for additional examples.

If Fisher's sharp null is true then $R_k = r_{Ck}$, and hence $T_k(Z, R_k) = T_k(Z, r_{Ck})$. For a particular $\Gamma > 1$, the test statistic's null distribution under Fisher's sharp null is

$$\mathrm{pr}\{T_k(Z, r_{Ck}) \geq v \mid \mathcal{F}, \mathcal{Z}\} = \sum_{z \in \Omega} 1\{T_k(Z, r_{Ck}) \geq v\} \frac{\exp\left(\gamma z^T u\right)}{\sum_{b \in \Omega} \exp\left(\gamma b^T u\right)}, \tag{3}$$

where $1(A)$ is an indicator that the condition $A$ was met. At $\Gamma = 1$ (3) is simply the proportion of treatment assignments where the test statistic is greater than or equal to $v$, returning the usual randomization inference in a finely stratified experiment. For $\Gamma > 1$ (3) is unknown due to its dependence on the nuisance vector $u$. A sensitivity analysis proceeds for a particular $\Gamma$ by maximizing (3) with $v = t_k$, the observed value of the test statistic for a particular $\Gamma$, resulting in the largest possible $p$-value for the desired inference subject to (1) holding at $\Gamma$. The practitioner then increases $\Gamma$ until the test no longer rejects the null hypothesis. This changepoint value of $\Gamma$ serves as a measure of how robust the study's finding was to unmeasured confounding. See Gastwirth et al. [2000] and Rosenbaum [2018] for large-sample approaches for conducting a sensitivity analysis for Fisher's sharp null with a single outcome variable under (1). Since $T_k$ is assumed effect increasing, the worst-case $p$-value for a sensitivity analysis for Fisher's sharp null attains the largest $p$-value over the composite null $H_k$. That is, a sensitivity analysis for Fisher's sharp null also provides a valid sensitivity analysis for $H_k$; see Proposition 1 of the unpublished manuscript of D. Caughey, A. Dafoe and L. Miratrix (arXiv:1709.07339) for a proof.

## 3 Sensitivity Analysis with multiple outcomes

### 3.1 A directional global null hypothesis

There are $K$ hypotheses $H_1, ..., H_K$, one each for the null of non-positive treatment effects for the $k$th outcome variable variable. We concern ourselves with a level-$\alpha$ sensitivity analysis for the global null hypothesis that all $K$ of these

hypotheses are true,

$$H_0 : \bigwedge_{k=1}^{K} H_k. \tag{4}$$

Through closed testing [Marcus et al., 1976], a valid sensitivity analysis for (4) also facilitates tests of the outcome-specific hypotheses $H_k$ while controlling the familywise error rate. See Fogarty and Small [2016, §5] for more on closed testing procedures applied to sensitivity analyses.

### 3.2 Linear combinations of test statistics and their distribution

In what follows it is useful to define $\varrho_{ij} = \mathrm{pr}\,(Z_{ij} = 1 \mid \mathcal{F}, \mathcal{Z})$. Under the global null (4) and recalling that our test statistics are of the form $T_k = Z^T q_k$ with $q_k$ fixed under the global null, the expectation $\mu(\varrho)$ and covariance $\Sigma(\varrho)$ for the vector of test statistics $T = (T_1, \ldots, T_K)^T$ are

$$\mu(\varrho)_k = \sum_{i=1}^{I} \sum_{j=1}^{n_i} q_{ijk} \varrho_{ij}, \quad \Sigma(\varrho)_{k,\ell} = \sum_{i=1}^{I} \left\{ \sum_{j=1}^{n_i} q_{ijk} q_{ij\ell} \varrho_{ij} - \left( \sum_{j=1}^{n_i} q_{ijk} \varrho_{ij} \right) \left( \sum_{j=1}^{n_i} q_{ij\ell} \varrho_{ij} \right) \right\}.$$

For a given vector of probabilities $\varrho$, under suitable conditions on the constants $q_{ijk}$ the distribution of $T$ is asymptotically multivariate normal through an application of the Cramér-Wold device. That is, for any fixed nonzero vector $\lambda = (\lambda_1, ..., \lambda_K)^T$ the standardized deviate $\lambda^T \{T - \mu(\varrho)\} / \{\lambda^T \Sigma(\varrho) \lambda\}^{1/2}$ is asymptotically standard normal.

The actual values of $\varrho$ are unknown to the practitioner due to their dependence on hidden bias. Instead, the constraints imposed by the sensitivity model (1) on $\varrho$ can be represented by an affine set. For a particular $\Gamma$ this set, call it $\mathcal{P}_\Gamma$, contains vectors $\varrho$ such that (i) $\varrho_{ij} \geq 0$ $(i = 1, ..., I; j = 1, ..., n_i)$; (ii) $\sum_{i=1}^{n_i} \varrho_{ij} = 1$ $(i = 1, ..., I)$; and (iii) $s_i \leq \varrho_{ij} \leq \Gamma s_i$ for some $s_i$ $(i = 1, ..., I; j = 1, ..., n_i)$. Conditions (i) and (ii) simply reflect that $\varrho_{ij}$ are probabilities, while the $s_i$ terms in (iii) arise from applying a Charnes-Cooper transformation [Charnes and Cooper, 1962] to (2).

### 3.3 Multivariate sensitivity analysis through a two-person game

Let $t = (t_1, ..., t_K)^T$ be the observed vector of test statistics. In this subsection only, suppose interest lies not in a test of (4), but rather in the narrower intersection null that Fisher's sharp null holds for all $K$ outcome variables. For fixed $\lambda$, a large-sample sensitivity analysis for Fisher's sharp null could be achieved by minimizing the standardized deviate $\lambda^T \{t - \mu(\varrho)\} / \{\lambda^T \Sigma(\varrho) \lambda\}^{1/2}$ over all $\varrho$ such that $\varrho \in \mathcal{P}_\Gamma$, and assessing whether the minimal objective value exceeds the appropriate critical value from a standard normal.

With $\lambda$ pre-specified, the sensitivity analysis imagines what would happen if the worst-case, adversarial bias at a given level of $\Gamma$ were present. If the practitioner fixes the linear combination $\lambda$ ahead of time, she has no further recourse against such adversarial attacks. The practitioner may instead consider a two-person game of the form

$$a_{\Gamma,\Lambda}^* = \min_{\varrho \in \mathcal{P}_\Gamma} \sup_{\lambda \in \Lambda} \frac{\lambda^T \{t - \mu(\varrho)\}}{\{\lambda^T \Sigma(\varrho) \lambda\}^{1/2}}, \tag{5}$$

where $\Lambda$ is some subset of $\mathbb{R}^K$ without the zero vector. The adversary may be thought of as embodying future counterclaims regarding the study's conclusions. In keeping with the scientific method the investigator recognizes that her conclusions will be subjected to challenges by her peers, and through the sensitivity analysis assesses whether a particular counterclaim could possibly overturn the study's findings. The critic aligns the unobserved confounders to inflate the $p$-value for the performed inference, while the investigator may choose weights for each outcome within the constraints imposed by $\Lambda$ in response to the configuration of unmeasured confounders selected by the critic. As $\Gamma$ increases, the critic has more flexibility with which to select potentially unfavorable treatment assignment distributions.

Most familiar large-sample sensitivity analyses for Fisher's sharp null hypothesis are instances of this game for particular choices of $\Lambda$. Setting $\Lambda = \{e_k\}$ where $e_k$ is a vector with a 1 in the $k$th coordinate and zeroes elsewhere returns a univariate sensitivity analysis for the $k$th outcome with a greater-than alternative, while $-e_k$ would return the less-than alternative. When the test statistics $T_K$ are rank tests, setting $\Lambda = \{1_K\}$ where $1_K$ is a vector containing $K$ ones returns the coherent rank test of Rosenbaum [1997]. When $\Lambda = \{e_1, ..., e_K\}$, the collection of standard basis vectors, (5) returns the method of Fogarty and Small [2016] with greater-than alternatives, and $\Lambda = \{\pm e_1, ..., \pm e_K\}$ gives the same method with two-sided alternatives. The method of Rosenbaum [2016] amounts to a choice of $\Lambda = \mathbb{R}^K \setminus \{0_K\}$, i.e. all possible linear combinations except the vector $0_K$ containing $K$ zeroes.

While appealing as a unifying framework for multivariate sensitivity analyses, the form (5) would be of little practical use if the corresponding optimization problem could not be readily solved. The problem (5) is not itself convex; however, consider replacing it with

$$b^*_{\Gamma,\Lambda} = \min_{\varrho \in \mathcal{P}_\Gamma} \; \sup_{\lambda \in \Lambda} \; \max\left[0, \frac{\lambda^T\{t - \mu(\varrho)\}}{\{\lambda^T \Sigma(\varrho)\lambda\}^{1/2}}\right]^2, \tag{6}$$

and let $f(\lambda, \varrho) = \max[0, \lambda^T\{t - \mu(\varrho)\}/\{\lambda^T\Sigma(\varrho)\lambda\}^{1/2}]^2$. This replaces negative values for the standardized deviate with zero, and then takes the square of the result. It is a monotone non-decreasing transformation of the standardized deviate in general, and is strictly increasing whenever the standardized deviate is larger than zero. The following proposition, proved in the Supplementary Material, establishes convexity of (6).

**Proposition 1.** *The function $g(\varrho) = \sup_{\lambda \in \Lambda} f(\lambda, \varrho)$ is convex in $\varrho$ for any set $\Lambda$ without the zero vector.*

The proof requires showing that for any $\lambda \in \Lambda$, the function $f(\lambda, \varrho)$ is convex in $\varrho$. As the pointwise supremum over a potentially infinite set of convex functions is itself convex [Boyd and Vandenberghe, 2004, §3.2.3], the result then follows. The convexity of $g(\varrho)$ allows for its minimization over the polyhedral set $\mathcal{P}_\Gamma$ such that the value $b^*_{\Gamma,\Lambda}$ in (6) can be computed in practice. For any $\Gamma$ and $\Lambda$, a sensitivity analysis through (5) would proceed by comparing the value $a^*_{\Gamma,\Lambda}$ to a suitable critical value $c_{\alpha,\Lambda}$. Observe that $a^*_{\Gamma,\Lambda} = (b^*_{\Gamma,\Lambda})^{1/2}$ for $a^*_{\Gamma,\Lambda} \geq 0$. If $\alpha \leq 0.5$ then $c_{\alpha,\Lambda}$ is non-negative for any choice of $\Lambda$. Consequently $a^*_{\Gamma,\Lambda} \geq c_{\alpha,\Lambda}$, leading to a rejection of the null, if and only if $b^*_{\Gamma,\Lambda} \geq c^2_{\alpha,\Lambda}$ so long as $\alpha \leq 0.5$. Through this equivalence, a large-sample sensitivity analysis using (5) can proceed through the solution of the convex program (6).

### 3.4 The practitioner's price

The critical value $c_{\alpha,\Lambda}$ depends on the structure of $\Lambda$, through which it is seen that additional flexibility in the set $\Lambda$ does not come without a cost. Intuition for the price to be paid may be formed at $\Gamma = 1$ in (5). When $\Lambda$ is a singleton the asymptotic reference distribution is the standard normal. If $\Lambda$ is instead a finite set with $|\Lambda| = L > 1$ simply comparing the optimal value of (4) to the $1 - \alpha$ quantile of a standard normal would not provide a level$-\alpha$ test due to multiplicity issues. One could proceed using a Bonferroni correction based on the $L$ comparisons, which would inflate the critical value. When $\Lambda = \mathbb{R}^K \setminus \{0_K\}$, Rosenbaum [2016] applies a result on quadratic forms of multivariate normals [e.g. Rao, 1973, page 60, 1f.1(i)] to show that one must instead use the square root of a critical value from a $\chi^2_K$ distribution when conducting inference through (5). This result underpins Scheffé's method for multiplicity control while comparing all linear contrasts of a multivariate normal (Scheffé, 1953). In the potential presence of hidden bias, the additional flexibility afforded by a richer set $\Lambda$ often offsets the loss in power from controlling for multiple comparisons, particularly in large samples. We discuss this further in §5.1, but see also Fogarty and Small (2016, §6) and Rosenbaum (2016, §4).

## 4 The Null Distribution Over Coherent Combinations

### 4.1 Adaptive linear combinations over the non-negative orthant

By allowing the set $\Lambda$ to be arbitrary, the developments §3.3 were presented with Fisher's sharp null in mind. A moment's reflection reveals that should inference instead concern the composite null (4) of non-positive effects for all outcome variables, the set $\Lambda$ must be constrained to maintain the desired size of the procedure. If $\Lambda$ allows for arbitrary linear combinations, evidence consistent with non-positive treatment effects for each outcome variable may nonetheless result in a rejection of the null hypothesis based on (5) beyond the nominal rate by setting $\lambda_k$ negative for each $k$. Directional control is lost without constraining the signs of the elements of $\Lambda$.

Following Rosenbaum (2002, §9.4), we define a family of coherent test statistics by restricting the vector $\lambda$ to lie in the non-negative orthant, $\Lambda_+ = \{\lambda : \lambda_k \geq 0 \; (k = 1..., K); \; \sum \lambda_k > 0\}$. The coherent test of Rosenbaum (1997) with $\lambda = 1_K$ is a particular element of $\Lambda_+$. We instead consider a large-sample sensitivity analysis for (5) with $\Lambda = \Lambda_+$, hence optimizing over the entire space of coherent linear combinations. We describe a projected subgradient descent method for solving (6) with $\Lambda = \Lambda_+$ in the Supplementary Material. Subgradients are straightforward to compute, and projections onto $\mathcal{P}_\Gamma$ are facilitated by the constraints being separable across matched sets.

Let $\varrho^*$ be the true, though typically unknown, vector of assignment probabilities and consider the random variable

$$A_{\Lambda_+}(Z, R) = \sup_{\lambda \in \Lambda_+} \frac{\lambda^T\{T - \mu(\varrho^*)\}}{\{\lambda^T \Sigma(\varrho^*)\lambda\}^{1/2}}. \tag{7}$$

Let $R_Z$ denote the observed responses when the treatment assignment is $Z$. Let $G(v, R_Z)$ be the reference distribution based on the observed outcome $R_Z$ assuming Fisher's sharp null,

$$G(v, R_Z) = \sum_{b \in \Omega} 1\{A_{\Lambda_+}(b, R_Z) \leq v\} \text{pr}(Z = b \mid \mathcal{F}, \mathcal{Z}), \tag{8}$$

and let $G^{-1}(1 - \alpha, R_Z)$ be its $1 - \alpha$ quantile. Observe that the reference distribution $G(v, R_Z)$ itself varies over elements of $\Omega$ through its dependence on $R_Z$ if Fisher's sharp null is false.

Proposition 2, proved in the Supplementary Material, states that a valid test of the composite null of non-positive effects $H_0$ can be achieved through the randomization distribution of $A_{\Lambda_+}$ under the assumption of Fisher's sharp null. Through an analogous proof, the randomization distribution also provides an unbiased test against positive alternatives of the form $\tau_{ijk} \geq 0$ $(i = 1, .., I; j = 1, ..., n_i; k = 1, ..., K)$ with at least one strict inequality.

**Proposition 2.** *Suppose that the global null (4) of non-positive treatment effects is true and assume that the test statistics $T_k$ $(k = 1, ..., K)$ are effect increasing. Then*

$$pr\{A_{\Lambda_+}(Z, R_Z) \geq G^{-1}(1 - \alpha, R_Z)\} \leq \alpha,$$

*such that the reference distribution under Fisher's sharp null controls the Type I error rate for any element of the composite null $H_0$.*

Both the observed value $A_{\Lambda_+} = a_{\Lambda_+}$ and the probabilities $\text{pr}(Z = z \mid \mathcal{F}, \mathcal{Z})$ are unknown in the observational study at hand due to their dependence on the true conditional assignment probabilities $\varrho^*$. Through the solution to (5) we instead observe the value $a^*_{\Gamma, \Lambda_+}$, which bounds $a_{\Lambda_+}$ from below so long as $\varrho^* \in \mathcal{P}_\Gamma$. That said, the true randomization distribution (8) typically remains unknown outside of a randomized experiment as it depends on $\varrho^*$. For many test statistics, such as those formed when $\Lambda$ is a singleton, the asymptotic reference distribution does not depend on $\varrho^*$ after suitable standardization. In what follows, we consider the large-sample distribution of $A_{\Lambda_+}$ under Fisher's sharp null.

## 4.2 The chi-bar-squared distribution

Comparing the optimal value of (5) with $\Lambda = \Lambda_+$ to the $1 - \alpha$ quantile of a standard normal would not provide a valid level$-\alpha$ sensitivity analysis, as it would not account for the optimization over coherent combinations. While one could proceed with the square root of the $1 - \alpha$ quantile of a $\chi^2_K$ distribution, doing so would be unduly conservative. The $\chi^2_K$ critical value allows for optimization over all linear combinations, while here we have constrained ourselves to combinations lying in the non-negative orthant. Theorem 1 provides the appropriate reference distribution given this restriction.

**Theorem 1.** *Suppose that $I^{-1}\Sigma(\varrho^*)$ has an positive definite limit $M$ as $I \to \infty$ and the random vector $\Sigma(\varrho^*)^{-1/2} \{T - \mu(\varrho^*)\}$ converges in distribution to a $K$-dimensional vector of independent standard normals. Then, as $I \to \infty$ the random variable $A^2_{\Lambda_+}$ converges in distribution to a $\bar{\chi}^2(M^{-1}, \Lambda_+)$ random variable under Fisher's sharp null.*

The proof is deferred to the Supplementary Material. The Supplementary Material also contains a discussion of sufficient conditions such that $\Sigma(\varrho^*)^{-1/2}\{T - \mu(\varrho^*)\}$ converges in distribution to a multivariate normal, which amount to assumptions about the vectors of constants $q_k$ $(k = 1, ..., K)$. For instance, one sufficient condition would be to stipulate that $I^{-1} \sum_{i=1}^I \sum_{j=1}^{n_i} q_{ijk}^4$ is uniformly bounded for all $I \in \mathbb{N}$ and all $k = 1, \ldots, K$.

The $\bar{\chi}^2$ ("chi-bar-squared") is a common family of distributions arising in order restricted statistical inference. To illustrate, let $X$ be a mean zero $K$-variate normal random vector with positive definite covariance matrix $V$, and define the random variable

$$\bar{\chi}^2(V, \Lambda_+) = X^T V^{-1} X - \inf_{\theta \in \Lambda_+} (X - \theta)^T V^{-1} (X - \theta). \tag{9}$$

Letting $\theta$ denote the mean vector of a multivariate normal, (9) is equivalent to the likelihood ratio statistic for testing the null $H_0 : \theta_k = 0$ $(k = 1, ..., K)$ versus the alternative $H_a : \theta_k \geq 0$ $(k = 1, ..., K)$ with strict inequality in at least one component (Kudô, 1963). Observe that replacing $\Lambda_+$ with $\mathbb{R}^K$ in (9) would return $X^T V^{-1} X$, and with it the usual $\chi^2_K$ distribution. Computation of (9) requires solving a quadratic program, an easy task with modern solvers but one which historically limited the adoption of methods requiring the $\bar{\chi}^2$ distribution.

The cumulative distribution function of the $\bar{\chi}^2(V, \Lambda_+)$ is $\text{pr}\{\bar{\chi}^2(V, \Lambda_+) \leq c\} = \sum_{i=0}^K w_i(V, \Lambda_+) \text{pr}\left(\chi^2_i \leq c\right)$, a mixture of $\chi^2_i$ distributions $(i = 0, ..., K)$ with $\chi^2_0$ representing a pointmass at zero. The $i$th weight $w_i(V, \Lambda_+)$ is equal to the probability that the vector $V^{-1/2}X$ has exactly $i$ positive components. The weights depend upon the covariance

$V$ through the corresponding correlation matrix $C$: any two covariance matrices $V'$ and $V$ with the same correlation structure $C$ yield the same weights for $\bar{\chi}^2$ [Silvapulle and Sen, 2005, Proposition 3.6.1 (11)]. See Kudô (1963), Robertson et al. (1988); and Silvapulle and Sen (2005) for more on the role of the $\bar{\chi}^2$ distribution in multivariate one-sided testing.

Shapiro (2003) presents an extension of Scheffé's method for multiple comparisons to linear combinations subject to cone constraints such as lying in the non-negative orthant. Arguments therein show that strong duality holds in (7), such that the optimal value for (7), $A_{\Lambda_+}$, equals the optimal value of the dual. The optimal solution to the dual is

$$A_{\Lambda_+} = \left\{ h^T \Sigma(\varrho^*) h - \inf_{\lambda \in \Lambda_+} (h - \lambda)^T \Sigma(\varrho^*)(h - \lambda) \right\}^{1/2}, \tag{10}$$

where $h = \Sigma^{-1}(\varrho^*)\{T - \mu(\varrho^*)\}$. Under mild conditions $h$ is asymptotically multivariate normal with covariance equal to the limit of $I\Sigma^{-1}(\varrho^*)$. Comparing (10) to (9) provides intuition for the $\bar{\chi}^2$ limiting distribution. Moving forwards, we refer to the procedure using $A_{\Lambda_+}$ to facilitate inference as the $\bar{\chi}^2$-test. In the Supplementary Material, we present Type I error control simulations indicating that the $\bar{\chi}^2$ reference distribution provides a reasonable approximation to the true randomization distribution of $A_{\Lambda_+}$ with moderate sample sizes.

### 4.3 The critical value and its dependence on the unknown assignment probabilities

A large-sample sensitivity analysis can be conducted by comparing the optimal value of (5) over coherent linear combinations, $a_{\Gamma,\Lambda_+}$ to the square root of the $1 - \alpha$ quantile of a $\bar{\chi}^2\{\Sigma^{-1}(\varrho^*), \Lambda_+\}$ distribution. Recalling that $\varrho^*$ is the true vector of assignment probabilities, we are faced with a difficulty encountered by neither a univariate sensitivity analysis for a particular outcome nor the method of Rosenbaum (2016): The asymptotic reference distribution depends on the assignment probabilities $\varrho^*$ through the covariance $\Sigma(\varrho^*)$ even after proper normalization. While $\varrho^*$ is known in a randomized experiment, the purpose of a sensitivity analysis is to assess robustness of a study's findings as $\varrho^*$ is allow to vary within bounds imposed by $\Gamma$.

The dependence of the covariance on nuisance parameters is commonly encountered in applications of the $\bar{\chi}^2$ distribution (Sen and Silvapulle, 2002, §2.2). One solution is to compute $p$-values through the bound $\mathrm{pr}\{\bar{\chi}^2(V, \Lambda_+) \geq c\} \leq 0.5\{\mathrm{pr}\left(\chi^2_{K-1} \geq c\right) + \mathrm{pr}\left(\chi^2_K \geq c\right)\}$; see Perlman (1969, Theorem 6.2) for a proof. This upper bound is attained in the limit as the correlation between all outcomes converges to one, and can itself be quite conservative in the presence of more moderate degrees of correlation typically observed in practical applications.

Motivated by the particular structure imposed by a sensitivity analysis, we instead use a two-stage procedure to better upper bound the worst-case critical value for each $\Gamma$. In a sensitivity analysis the range of the nuisance parameters $\varrho^*$ is controlled by $\Gamma$. At $\Gamma = 1$ $\varrho^*$ is entirely specified, such that in finely stratified experiments the appropriate $\bar{\chi}^2$ distribution is known. As $\Gamma$ increases the bounds imposed by membership in $\mathcal{P}_\Gamma$ widen. For each pair of outcomes $k$ and $\ell$, we first find upper and lower bounds on the correlation between $k$ and $k'$ given $\varrho^* \in \mathcal{P}_\Gamma$, call them $C^{(\ell)}_{k,k',\Gamma}$ and $C^{(u)}_{k,k',\Gamma}$. We then maximize the $1 - \alpha$ quantile of a $\bar{\chi}^2(C^{-1}, \Lambda_+)$ distribution over the correlation matrix $C$ subject to $C^{(\ell)}_{k,k',\Gamma} \leq C_{k,k'} \leq C^{(u)}_{k,k',\Gamma}$ for all $k, k'$ and $C$ being a correlation matrix. See the Supplementary Material for implementation details along with a discussion of the case $K = 2$, where it is seen that the worst-case critical value is attained at the lower bound on the correlation. In practice, we find that this can provide meaningful improvements in the power of the procedure; see the Supplementary Material for an illustration.

## 5 Design Sensitivity and Power for the $\bar{\chi}^2$-Test

### 5.1 Design sensitivity

Suppose that the treatment in question actually has an effect in the direction of the alternative, and further that there is truly no hidden bias such that inference at $\Gamma = 1$ would be justified. As would be the case in practice, the researcher analyzing the observational study is unaware of these favorable conditions. Thus, she would like to reject the null hypothesis not only under the assumption of no unmeasured confounding, but also for values $\Gamma > 1$ to assess whether the rejection of the null is robust to certain degrees of hidden bias. The power of a level-$\alpha$ sensitivity analysis is the probability that the procedure correctly rejects the null hypothesis at some pre-specified value of $\Gamma \geq 1$. In what follows we will assume a stochastic generative model for the outcome variables, an assumption which greatly simplifies power calculations.

Under mild conditions, there is a value $\tilde{\Gamma}$ such that the power of a sensitivity analysis converges to one for all $\Gamma < \tilde{\Gamma}$, and converges to zero for all for all $\Gamma > \tilde{\Gamma}$; this value is called the design sensitivity of the test (Rosenbaum, 2004). It

| | $\bar{\chi}^2$-Test | | Equal-Weight Test | | Max Univariate | |
|---|---|---|---|---|---|---|
| | $\rho = 0$ | $\rho = 0.2$ | $\rho = 0$ | $\rho = 0.2$ | $\rho = 0$ | $\rho = 0.2$ |
| $\tau = (0.25, 0.25, 0.25)^T$ | 2.9 | 2.4 | 2.9 | 2.4 | 1.9 | 1.9 |
| $\tau = (0.10, 0.10, 0.50)^T$ | 3.6 | 3.4 | 2.6 | 2.2 | 3.4 | 3.4 |
| $\tau = (0.02, 0.20, 0.50)^T$ | 3.8 | 3.5 | 2.8 | 2.4 | 3.4 | 3.4 |

Table 1: Design sensitivities for $\bar{\chi}^2$-test, the equal-weight test, and the largest of the three univariate tests under both independence and and moderate positive correlation between outcomes.

quantifies the asymptotic ability of the test to discriminate treatment effect under the concern of bias in the treatment allocation process, and can vary substantially across choices of test statistics. For a fixed data generating model, a test with high design sensitivity is preferable to a test with low design sensitivity.

For fixed choices of the univariate test statistics $T_k = Z^T q_k$ $(k = 1, ..., K)$, we consider the design sensitivity of multivariate tests based upon (5) and their dependence on the set $\Lambda$. Theorem 2 shows that design sensitivity is a monotonic non-decreasing function with respect to the partial ordering over sets $\Lambda$ given by inclusion.

**Theorem 2.** *Suppose $\Lambda_1 \subseteq \Lambda_2$. Under mild conditions, the design sensitivity of (5) using $\Lambda = \Lambda_1$ is less than or equal to the design sensitivity of (5) using $\Lambda = \Lambda_2$.*

The proof of Theorem 2 is deferred to the Supplementary Material. In light of Theorem 2, it may be tempting to take $\Lambda = \mathbb{R}^K \setminus \{0_K\}$ in order to achieve the greatest design sensitivity. While this would result in a valid test of Fisher's sharp null, it does not provide a valid test of the null hypothesis of non-positive treatment effects: should the signs of $\lambda_k$ be left unconstrained, evidence of a negative treatment effect may result in a large optimal value for (5). Restricting attention to the set of coherent linear combinations $\Lambda_+$, Theorem 2 gives rise to the following optimality property for the $\bar{\chi}^2$-test due to its optimizing over the entirety of $\Lambda_+$.

**Corollary 1.** *The $\bar{\chi}^2$-test achieves greatest design sensitivity among coherent tests based upon (5) with $\Lambda \subseteq \Lambda_+$*

### 5.2 Finite-sample power for rejecting the global null

Corollary 1 illustrates that despite the larger critical value necessitated by the $\bar{\chi}^2$-test by optimizing over $\Lambda_+$, the $\bar{\chi}^2$-tests achieves the largest possible design sensitivity over the set of coherent multivariate tests. This reflects that in large samples bias trumps variance in the analysis of observational studies, such that the differences in critical values are rendered irrelevant in the limit. In moderate samples, the variance of the null distribution plays a larger role in the power of a sensitivity analysis, such that differences in critical values can make a more substantial difference for procedures with similar design sensitivities.

We present a simulation study comparing the power of a sensitivity analysis based upon the $\bar{\chi}^2$-test to two competitors: the method of Fogarty and Small (2016); and the test using (5) with $\Lambda = \{1_K\}$, which we refer to as the equal-weight test. Combining test statistics with equal weights is only sensible when the constituent test statistics $T_k$ $(k = 1, .., K)$ reflect evidence against the null hypothesis on the same scale. This would be true of rank statistics as described in Rosenbaum (1997), and would also be true of suitably scaled $m$-statistics of the type described in Rosenbaum (2007); however, if one outcome is tested using a rank-sum statistic and another with an $m$-statistic for instance, the "equal-weight" test would give unreasonable weight to the rank-sum recorded outcome. The $\bar{\chi}^2$-test and the test of Fogarty and Small (2016) do not require comparable scales for the test statistics as they are scale invariant.

The simulations are performed on $I = 300$ matched pairs with $K = 3$ outcomes. In each simulation, we generate $I$ mean-zero unit-variance trivariate normal vectors of noise $(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})^T$ equicorrelated with correlation $\rho$. We then create the vector of treated-minus-control paired differences in outcomes as $(Y_{i1}, Y_{i2}, Y_{i3})^T = (\tau_1, \tau_2, \tau_3)^T + (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})^T$ for different values of the treatment effects $(\tau_1, \tau_2, \tau_3)^T$. For each outcome variable, the employed test statistic is $T_k = \sum_{i=1}^I \text{sign}(Y_{ik}) \min(|Y_{ik}|/s_k, 2.5)$, where $s_k$ is the median of $|Y_{ik}|$ $(i = 1, .., I)$. This amounts to a choice of a $m$-statistic with Huber's $\psi$-function, as described in Rosenbaum (2007).

Table 1 presents the values of the treatment effects and the correlation employed in the simulation study. For each combination of parameters, it further provides the design sensitivity for the $\bar{\chi}^2$-test and the equal-weight test. While there is no known formula for the design sensitivity of the procedure of Fogarty and Small (2016), it is lower-bounded by the the maximal design sensitivities of the three univariate tests; this value is also presented in the table. The table reflects Corollary 1: for each combination of parameters, the design sensitivity for the $\bar{\chi}^2$-test is greater than or equal to that of the equal-weight test and the maximal univariate test. Further, there is no consistent ordering between the equal-weight test and the max of the univariate tests, as the corresponding sets $\Lambda$ for neither test is a subset of the other.
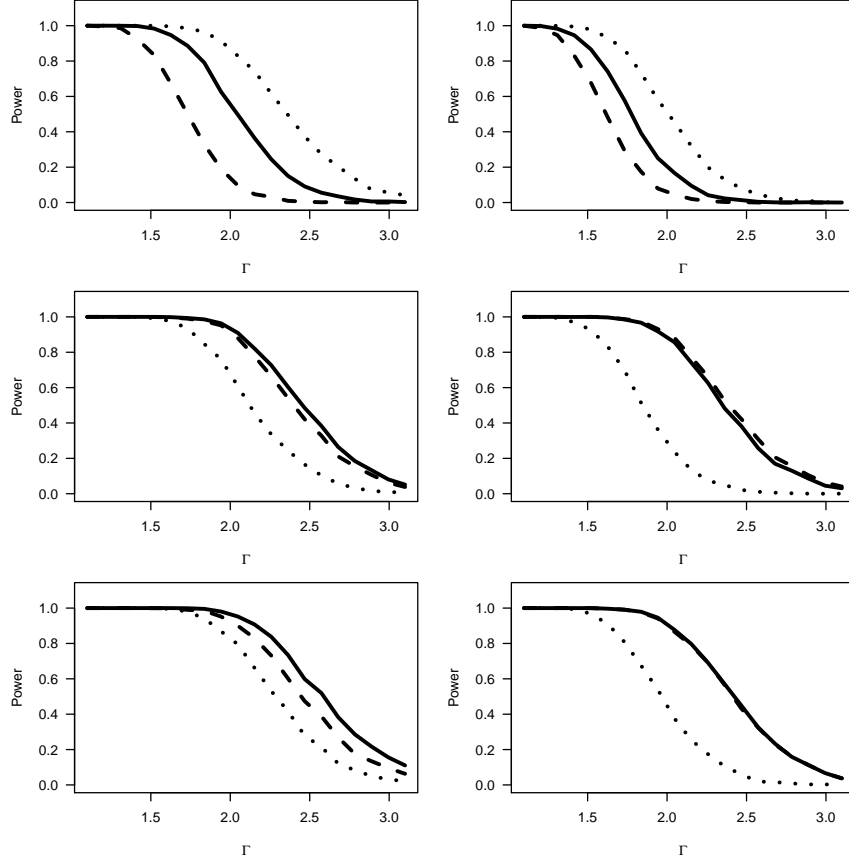
8

Figure 1: Power comparisons between the method of Fogarty and Small (2016) (dashed), the $\bar{\chi}^2$-test of this paper (solid), and the equal-weight test (dotted) as $\Gamma$ increases with $I = 300$. The first row has $\tau_1 = \tau_2 = \tau_3 = 0.25$; the second row has $\tau_1 = \tau_2 = 0.1$ and $\tau_3 = 0.5$; and the third row has $\tau_1 = 0.05$, $\tau_2 = 0.2$, and $\tau_3 = 0.5$. The first column has $\rho = 0$, while the second has $\rho = 0.2$.

Figure 1 presents the estimated power curves of the three tests as a function of $\Gamma > 1$ in these simulation settings at $I = 300$, with 2000 simulations for each combination of parameters. The correlation between paired differences varies across the columns from $\rho = 0$ (left) to $\rho = 0.2$ (right), while the treatment effects vary down the rows. The first row corresponds to $\tau_1 = \tau_2 = \tau_3$ and $I$, and here it is seen that the equal-weight test outperforms both the $\bar{\chi}^2$-test and Fogarty and Small (2016). When the treatment effects are equal the linear combination $\lambda = 1_K$ attains the largest design sensitivity, and by restricting $\Lambda$ to only this linear combination the lower critical value employed by the equal-weighted test improves power over that attained by the $\bar{\chi}^2$-test. When one of the three outcomes is strongly affected by the treatment while the other two are minimally impacted, as in the second row of the figure, the method of Fogarty and Small (2016) and the $\bar{\chi}^2$-test perform similarly, while the equal-weight test lags behind. The test statistics returned by the $\bar{\chi}^2$-test are larger, but this is offset relative to the method of Fogarty and Small (2016) by the larger critical value necessitated. When the treatment effects are staggered between the three outcomes as in the third row, the $\bar{\chi}^2$-test outperforms both Fogarty and Small (2016) and the equal-weight test, particularly in the case of independence between the outcome variables. Optimizing over $\Lambda_+$ increases the value of the test statistic over both competitors, such that the flexibility is well worth the price of a larger critical value.

The simulations indicate that while the $\bar{\chi}^2$-test must have optimal power in the limit as asserted by Corollary 1, it need not have the best finite-sample performance. In some cases the equal-weight test can outperform it, while in others it is outperformed by the method of Fogarty and Small (2016). Importantly the $\bar{\chi}^2$-test was never the worst of the three methods considered, and the simulations show that *a priori* restricting the set of combinations $\Lambda$ under consideration can substantially reduce power should the choice of $\Lambda$ be poor. The $\bar{\chi}^2$-test does pay a price in terms of an increased critical value, but this price acts as insurance against an unwise choice of $\Lambda$.

In the Supplementary Material, we present additional simulations with $I = 1000$ matched pairs which begin to show convergence of behavior of the tests under comparison to their design sensitivities. We further illustrate the potential for improvements in power for testing outcome-specific null hypotheses through incorporating the $\bar{\chi}^2$-test into a closed testing framework, as described in Fogarty and Small (2016, §6).

## 6 Illustrations of Multivariate One-Sided Sensitivity Analysis

### 6.1 The role of coherence in two observational studies

We now consider the role of multiple outcomes in two observational studies. Both examples are drawn from The National Health and Nutrition Examination Survey (NHANES) and study physiological impacts of cigarette smoking. One study investigates the impact of smoking on two measures of periodontal disease, while the other looks at whether smoking increases urinary metabolite levels of four carcinogens. In both examples, the alternative hypothesis is that smoking should have a positive treatment effect on each of the outcome variables measured. Rosenbaum remarks that "If incoherence presents a substantial obstacle to a claim that the treatment caused its ostensible effects, then the absence of incoherence - that is, coherence - should entail some strengthening of that claim" (Rosenbaum, 2010b, p. 119). Should the evidence suggest ostensible effects of smoking incompatible with positive effects for each outcome variable, smoking's place in the causal pathway would be cast into doubt. Should the outcomes all be affected in the predicted direction, this would provide further evidence for smoking's role in the causal mechanism.

In both observational studies and for each outcome variable, we use an $m$-test based upon Huber's $\psi$-function to conduct inference with the default choices for parameters in the `senmv` function in the `sensitivitymv` package in R.

### 6.2 Smoking and periodontal disease

It has been suggested that up to 42% of cases of periodontal disease can be attributed to smoking (Tomar and Asma, 2000); however, as the evidence is observational in nature this association may well be explained away by other intrinsic differences between smokers and non-smokers. Using the 2011-2012 NHANES survey, Rosenbaum (2016) paired $I = 441$ smoking individuals to non-smokers who were similar on the basis of education, income, race, age and gender. Two outcome variables pertaining to dental health were recorded, one each for upper and lower teeth. In this context, coherence would amount to demonstrating that smoking negatively impacted dental health in both the upper and lower teeth. Such a coherent hypothesis strengthens the causal claim that cigarette smoking is detrimental to periodontal health. Should smoking only appear to impact upper teeth but not lower teeth, for instance, such incoherence would cast into doubt whether smoking is truly to blame.

At $\alpha = 0.05$, the overall null hypothesis of non-positive treatment effects was rejected up until $\Gamma = 2.36$ when using the $\bar{\chi}^2$-test, while the equal-weight test was able to reject until $\Gamma = 2.54$. By selecting $\Lambda = \Lambda_+$ Theorem 1 gives that the appropriate asymptotic null distribution is the $\bar{\chi}^2$ distribution, while for the equal-weight test the asymptotic null distribution is the standard normal. The $1 - \alpha$ quantile of the standard normal lies below the square root of the $1 - \alpha$ quantile of any $\bar{\chi}^2$ distribution, such that the equal-weight test is able to employ a smaller critical value. With this restriction comes the risk that equally weighting the outcome variables may be suboptimal. In this particular observational study, it comes as little surprise that with periodontal disease in upper and lower teeth the risk was worth the while: there is little reason to suspect that magnitude of effects on upper and lower teeth should differ. Sensitivity analysis using the method of Fogarty and Small (2016) achieves significance up to $\Gamma = 2.32$, a slightly lower value than the $\bar{\chi}^2$-test. The method of Fogarty and Small (2016) takes $\Lambda$ as the set of standard unit basis vectors for $\mathbb{R}^2$ in this case, and does not combine the two related measures of periodontal disease. Despite the method of Fogarty and Small (2016) also having a smaller critical value than the $\bar{\chi}^2$-test, in this example this was offset by the additional flexibility afforded by the $\bar{\chi}^2$-test in optimizing over $\Lambda_+$. The sensitivity analysis using the $\bar{\chi}^2$-test took 13 seconds to complete on a personal laptop with a 2.60GHz processor with 16GB of RAM for this data set.

### 6.3 Smoking and polycyclic aromatic hydrocarbons

Polycyclic Aromatic Hydrocarbons (PAHs) are a class of organic compounds formed during incomplete combustion which have been labeled potentially carcinogenic to humans (Boström et al., 2002). We examine urinary concentrations of four different PAH metabolites in 432 smokers and 1206 non-smokers recorded in NHANES 2007-2008. The four metabolites are 1-hydroxyphenanthrene (1-Phen), 3-hydroxyphenanthrene (3-Phen), 1-hydroxypyrene (1-Pyr), and 9-hydroxyfluorene (9-Fluo). Full matching (Hansen, 2004) was employed to adjust for a host of measured covariates thought to impact one's decision to smoke and one's exposure to PAHs; see the supplementary material for additional details. We then proceed with inference assessing whether cigarette use increases urinary concentrations of these four PAH metabolites. As tobacco smoke contains all of these PAHs, an incoherent result that none or only some urinary

concentrations of PAH metabolites are higher in smokers than in non-smokers be discovered would cast into question whether the association between smoking cigarettes and urinary PAH concentrations was actually causal. At $\alpha = 0.05$, a sensitivity analysis using the $\bar{\chi}^2$-test yielded significance up to $\Gamma = 6.28$, whereas for the equal-weight test with $\Lambda = \{1_K\}$ the sensitivity analysis was only able to reject up to $\Gamma = 5.38$. Despite the smaller critical value, in this case restricting oneself to equal weighting led to a markedly lower changepoint value of $\Gamma$ than did the $\bar{\chi}^2$-test. The method of Fogarty and Small (2016) rejected until $\Gamma = 6.18$.

At $\Gamma = 6.28$, our procedure for upper bounding the worst-case critical value for the $\bar{\chi}^2$-test as described in §4.3 returns a bound of 2.20 for the test based upon $a^*_{6.18, \Lambda_+}$ in (5). To illustrate the improvements from this approach, the square root of the 0.95 quantile of a $\chi^2_4$ is 3.08, while employing the conservative bound from Perlman (1969, Theorem 6.2) yields a critical value of 2.96. The $\bar{\chi}^2$ sensitivity analysis ran in about 20 minutes on a personal laptop with a 2.60GHz processor with 16GB of RAM. The length of runtime is dependent upon several factors including the number of strata, the size of the strata, the number of outcome variables, and the number of values of $\Gamma$ tested in the sensitivity analysis.

### 6.4 Improvements in tests of individual null hypotheses

Rejecting the global null hypothesis confirms to the experimenter that at least one of the outcome variables is impacted by treatment in the direction of the alternative. However in order to appraise a coherent pattern of treatment impact an experimenter will need to examine the local null hypotheses of treatment impact upon each of the outcomes individually. Correcting for multiple comparisons can be facilitated through many techniques; here we juxtapose embedding the $\bar{\chi}^2$-test into a closed testing framework against performing $K$ individual sensitivity analyses, one for each outcome variable, while employing a Bonferroni correction.

|  | Periodontal Disease | | Polycyclic Aromatic Hydrocarbons | | | |
|---|---|---|---|---|---|---|
|  | Lower Teeth | Upper Teeth | 1-Phen | 3-Phen | 1-Pyr | 9-Fluo |
| Closed Testing | 2.26 | 1.82 | 2.13 | 5.28 | 5.25 | 5.78 |
| Bonferroni | 2.17 | 1.76 | 1.99 | 4.88 | 4.84 | 5.31 |
| Uncorrected | 2.26 | 1.82 | 2.13 | 5.28 | 5.25 | 5.78 |

Table 2: Comparison of the closed test changepoint $\Gamma$ versus Bonferroni corrected sensitivity analysis changepoint $\Gamma$ for the data examples at $\alpha = 0.05$. The last row is the benchmark given by conducting individual tests at $\alpha$ without correction for multiplicity.

Table 2 details the changepoint $\Gamma$ values for each individual outcome of the periodontal data and the PAH data while controlling the familywise error rate at $\alpha = 0.05$. The $\bar{\chi}^2$-test embedded into a closed testing framework outperformed the Bonferroni corrected tests for each outcome. Table 2 also includes the changepoint $\Gamma$ values returned by the univariate sensitivity analyses without a Bonferroni correction, i.e. with each outcome tested at $\alpha = 0.05$. The table reveals that through embedding the $\bar{\chi}^2$-test in a closed testing procedure, in both studies we are able to report the same robustness to unmeasured confounding that would have been attained had we not controlled for multiple comparisons in the first place. Due to the improvements in power along the closed testing path furnished by the $\bar{\chi}^2$-test, there is no cost for evaluating coherence of all outcome variables relative to the best univariate outcome analysis.

## Supplementary Material

Supplementary Material available contains theoretical results, simulation studies, further algorithmic details, additional insight into the $\bar{\chi}^2$ distribution, further information on the observational study on smoking and polycyclic aromatic hydrocarbons, and an R script for implementing the method proposed in this work.

# Supplementary Material

## A.1 Proof of Main Results

### A.1.1 Proposition 1

**Proposition 1.** *The function $g(\varrho) = \sup_{\lambda \in \Lambda} f(\lambda, \varrho)$ is convex in $\varrho$ for any set $\Lambda$ without the zero vector.*

In order to show that (6) is convex in $\varrho$ we first prove a lemma.

**Lemma 1.** *For a fixed $\lambda \in \mathbb{R}^K$ the function $d(\varrho) = \lambda^T \Sigma(\varrho) \lambda$ is a concave function of $\varrho$.*

*of Lemma 1.* Define $Q_i$ to be the $K$-by-$n_i$ matrix where the $(k, j)$th entry is $q_{ijk}$. Then the Hessian matrix of $d(\varrho)$ with respect to the variables in the $i$th strata is

$$\nabla^2_{\varrho_{ij};\ (j=1,...,n_i)} d(\varrho) = \frac{-1}{2} Q_i^T \lambda \lambda^T Q_i.$$

This is negative semi-definite. By independence between strata, the full Hessian $\nabla^2_\varrho f(\varrho)$ is the direct sum of the Hessians associated to each stratum. Thus, the full Hessian matrix is a block diagonal matrix wherein each block is negative semi-definite. Since the eigenvalues of a block diagonal matrix are the collection of eigenvalues of its constituent blocks, we have that the full Hessian must be negative semi-definite as well. As a consequence, $d(\varrho)$ is a concave function of $\varrho$. $\square$

*of Proposition 1.* The identity function $x \mapsto x$ is convex as a function of $x$. Since the point-wise maximum of convex functions is convex $\max\{0, x\}$ is convex as a function of $x$. The quadratic function $a \mapsto a^2$ is convex and increasing on the non-negative real line so by Boyd and Vandenberghe (2004, 3.10) the function $\psi(x) = [\max\{0, x\}]^2$ is a convex function of $x$.

The perspective of a function $\psi(x)$ is defined to be $\phi(x, v) = v\psi(x/v)$ for $v > 0$; by Boyd and Vandenberghe (2004, 3.2.6) the perspective of a convex function is convex as well. Computing the perspective of $\psi$ follows as

$$\begin{aligned}
\phi(x, v) &= v\psi(x/v) \\
&= v \left\{ \max(0, x/v) \right\}^2 \\
&= v \left\{ \frac{\max(0, x)}{v} \right\}^2 \\
&= \frac{\max(0, x)^2}{v}.
\end{aligned}$$

Thus, $\phi(x, v) = \max(0, x)^2/v$ is convex. Now, consider any fixed $\lambda \geq 0$ and $t$ of dimension $K$. $\mu(\varrho)$ is a linear function of $\varrho$. Since affine transformations of linear functions are convex, $\lambda^T \{t - \mu(\varrho)\}$ is convex. Furthermore, $\lambda^T \Sigma(\varrho) \lambda$ is concave in $\varrho$ by Lemma 1. By Boyd and Vandenberghe (2004, 3.15), since $\phi(x, v)$ is non-decreasing in $x$ and non-increasing in $v$ the function

$$f(\lambda, \varrho) = \phi\left(\lambda^T \{t - \mu(\varrho)\}, \lambda^T \Sigma(\varrho)\lambda\right) = \frac{\max\left[0, \lambda^T \{t - \mu(\varrho)\}\right]^2}{\lambda^T \Sigma(\varrho) \lambda}$$

is convex in $\varrho$. As $g(\varrho)$ is the point-wise supremum over all $\lambda \in \Lambda$ of $f(\lambda, \varrho)$, by Boyd and Vandenberghe (2004, 3.7) $g(\varrho)$ is convex in $\varrho$ as desired. The requirement that $\Lambda$ excludes the zero vector ensures that for any positive definite $\Sigma(\varrho)$ the denominator is always defined and thus $g(\varrho)$ is defined. $\square$

### A.1.2 Proposition 2

Here and elsewhere in the supplement, $\Lambda_+$ is once again defined to be the non-negative orthant in $\mathbb{R}^K$ excluding the zero vector, that is $\Lambda_+ = \{\lambda : \lambda_k \geq 0 \ \ (k = 1..., K);\ \ \sum \lambda_k > 0\}$.

**Proposition 2.** *Suppose that the global null (4) of non-positive treatment effects is true and assume that the test statistics $T_k\ (k = 1, ..., K)$ are effect increasing. Then*

$$pr\{A_{\Lambda_+}(Z, R_Z) \geq G^{-1}(1 - \alpha, R_Z)\} \leq \alpha,$$

*such that the reference distribution under Fisher's sharp null controls the Type I error rate for any element of the composite null $H_0$.*

*Proof.*

$$
\begin{aligned}
&\mathrm{pr}\{A_{\Lambda_+}(Z, R_Z) > G^{-1}(1-\alpha, R_Z)\} \\
&= \sum_{z \in \Omega} 1\{A_{\Lambda_+}(z, R_z) > G^{-1}(1-\alpha, R_z)\}\mathrm{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) \\
&= \sum_{b \in \Omega} \left[ \sum_{z \in \Omega} 1\{A_{\Lambda_+}(z, R_z) > G^{-1}(1-\alpha, R_z)\}\mathrm{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) \right] \mathrm{pr}(Z = b \mid \mathcal{F}, \mathcal{Z}) \\
&\leq \sum_{b \in \Omega} \left[ \sum_{z \in \Omega} 1\{A_{\Lambda_+}(b, R_z) > G^{-1}(1-\alpha, R_z)\}\mathrm{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) \right] \mathrm{pr}(Z = b \mid \mathcal{F}, \mathcal{Z}) \\
&= \sum_{z \in \Omega} \left[ \sum_{b \in \Omega} 1\{A_{\Lambda_+}(b, R_z) > G^{-1}(1-\alpha, R_z)\}\mathrm{pr}(Z = b \mid \mathcal{F}, \mathcal{Z}) \right] \mathrm{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) \\
&\leq \alpha \sum_{z \in \Omega} \mathrm{pr}(Z = z \mid \mathcal{F}, \mathcal{Z}) = \alpha.
\end{aligned}
$$

The third line simply multiplies by one in the form of $\sum_{b \in \Omega} \mathrm{pr}(Z = b \mid \mathcal{F}, \mathcal{Z})$. The fourth line uses that the test statistics are effect increasing. After rearranging the order of summation in the fifth line, the sixth follows by definition as it simply uses that for any particular $z$, $G^{-1}(1-\alpha, R_z)$ is the $1-\alpha$ quantile corresponding to $G(v, R_z) = \sum_{b \in \Omega} 1\{A_{\Lambda_+}(b, R_z) \leq v\}\mathrm{pr}(Z = b \mid \mathcal{F}, \mathcal{Z})$.

<div align="right">□</div>

### A.1.3   Theorem 1

For ease of notation we suppress conditioning on $\mathcal{F}$ and $\mathcal{Z}$ when writing expectations and covariances in this section. We again define $T_k = \sum_{i=1}^{I} \sum_{j=1}^{n_i} Z_{ij} q_{ijk}$, and let $\varrho^*$ represent the true vector of conditional assignment probabilities. For precision quantities such as $\varrho^*$ should be subscripted by $I$ to denote their dependence on the sample size; this is omitted for improved readability.

**Theorem 3.** *Suppose that $I^{-1}\Sigma(\varrho^*)$ has a positive definite limit $M$ as $I \to \infty$ and the random vector $\Sigma(\varrho^*)^{-1/2}\{T - \mu(\varrho^*)\}$ converges in distribution to a $K$-dimensional vector of independent standard normals. Then, as $I \to \infty$ the random variable $A_{\Lambda_+}^2$ converges in distribution to a $\bar{\chi}^2(M^{-1}, \Lambda_+)$ random variable under Fisher's sharp null.*

Before proving Theorem 1, we establish conditions under which the random vector $\Sigma(\varrho^*)^{-1/2}\{T - \mu(\varrho^*)\}$ has a multivariate normal limiting distribution.

**Lemma 2.** *Suppose that there exists a $\delta > 0$ for which*

$$
\sum_{i=1}^{I} E\left( \mid \sum_{j=1}^{n_i} q_{ijk} Z_{ij} - \sum_{j=1}^{n_i} q_{ijk}\varrho_{ij}^* \mid^{2+\delta} \right) = O(I) \tag{A.1}
$$

*for all $k$ and all $I$, and that $I^{-1}\Sigma(\varrho^*)$ has an positive definite limit $M$ as $I \to \infty$. Then as $I \to \infty$ the random vector $\Sigma(\varrho^*)^{-1/2}\{T - \mu(\varrho^*)\}$ converges in distribution to a $K$-variate vector of independent standard normals.*

*of Lemma 2.* Define $X_i = (X_{i1}, \ldots, X_{iK})^T$ where $X_{ik} = \sum_{j=1}^{n_i} q_{ijk} Z_{ij}$. Denote $\mu_i(\varrho^*) = E(X_i)$ and $\Sigma_i(\varrho^*) = E(X_i X_i^T) - E(X_i)E(X_i)^T$, such that $\sum_{i=1}^{I} \mu_i(\varrho^*) = \mu(\varrho^*) = E(T)$ and $\sum_{i=1}^{I} \Sigma_i(\varrho^*) = \Sigma(\varrho^*) = \mathrm{cov}(T)$.

By the Cramér-Wold device it suffices to consider the distribution of the univariate random variable $I^{-1/2}\sum_{i=1}^{I} \lambda^T\{X_i - \mu_i(\varrho^*)\}$ for a fixed, non-zero, $\lambda \in \mathbb{R}^K$. By independence between strata, the random variables $\lambda^T\{X_i - \mu_i(\varrho^*)\}$ are independent but not necessarily identically distributed. The variance of $I^{-1/2}\sum_{i=1}^{I} \lambda^T\{X_i - \mu_i(\varrho^*)\}$ is $I^{-1}\sum_{i=1}^{I} \lambda^T\Sigma_i(\varrho^*)\lambda$. By hypothesis $I^{-1}\Sigma(\varrho^*)$ has an positive definite limit $M$ as $I \to \infty$ so

$$
\lim_{I \to \infty} \frac{1}{\left( I^{-1}\sum_{i=1}^{I} \lambda^T\Sigma_i(\varrho^*)\lambda \right)^{\frac{2+\delta}{2}}} = \frac{1}{(\lambda^T M \lambda)^{\frac{2+\delta}{2}}} > 0. \tag{A.2}
$$

<div align="center">13</div>

Furthermore, (A.1) and the $c_r$-inequality imply that

$$\lim_{I \to \infty} I^{-\frac{2+\delta}{2}} \sum_{i=1}^{I} E\left( \mid \sum_{j=1}^{i} q_{ijk} Z_{ij} - \sum_{j=1}^{i} q_{ijk} \varrho_{ij} \mid^{2+\delta} \right) = 0. \tag{A.3}$$

Combining (A.2) and (A.3) gives that

$$\lim_{I \to \infty} \frac{1}{\left( \sum_{i=1}^{I} \lambda^T \Sigma_i(\varrho^*) \lambda \right)^{\frac{2+\delta}{2}}} \sum_{i=1}^{I} E\left( \mid \sum_{j=1}^{i} q_{ijk} Z_{ij} - \sum_{j=1}^{i} q_{ijk} \varrho_{ij} \mid^{2+\delta} \right) = 0.$$

The Lyapunov central limit theorem then implies that

$$\frac{\sum_{i=1}^{I} \lambda^T \{ X_i - \mu_i(\varrho^*) \}}{\left\{ \sum_{i=1}^{I} \lambda^T \Sigma_i(\varrho^*) \lambda \right\}^{1/2}}$$

converges in distribution to the standard univariate normal. Hence, the Cramér-Wold device establishes that $\Sigma(\varrho^*)^{-1/2} \{ T - \mu(\varrho^*) \}$ converges in distribution to a $K$-variate vector of independent standard normals. $\square$

The sufficient criterion given in the main text, that $I^{-1} \sum_{i=1}^{I} \sum_{j=1}^{n_i} q_{ijk}^4$ is uniformly bounded for all $I$ and all $k = 1, \dots, K$, satisfies the conditions of Lemma 2 with $\delta = 2$ since $Z_{ij}$ is binary and $0 \le \varrho_{ij}^* \le 1$ for all $i$ and $j$.

For many statistics, such as an $m$-statistic using Huber's $\psi$ function, $q_{ijk}$ are bounded for all $i, j$ and $k$. In these cases, asymptotic normality would hold if the stratum sizes $n_i$ were bounded, for instance. When the underlying $q_{ijk}$ varies as a function of $I$ as with various rank tests, the proof given above is insufficient. In such cases, a triangular array version of the central limit theorem must be applied and the sufficient conditions adapted accordingly to guarantee asymptotic normality as $I \to \infty$.

*of Theorem 1.* Consider the random variable

$$D^2_{\Lambda_+} = h^T \Sigma(\varrho^*) h - \inf_{\lambda \in \Lambda_+} (h - \lambda)^T \Sigma(\varrho^*)(h - \lambda), \tag{A.4}$$

where $h = \Sigma(\varrho^*)^{-1} \{ T - \mu(\varrho^*) \}$. Assume no degeneracy between the test statistics, such that the covariance matrix $\Sigma(\varrho^*)$ is positive definite for all $I$. For $\Sigma(\varrho^*)$ positive definite, the program

$$\inf_{\lambda \in \Lambda_+} (h - \lambda)^T \Sigma(\varrho^*)(h - \lambda) \tag{A.5}$$

is convex. Since the feasible region of (A.5) is $\Lambda_+$, the relative interior of the feasible region is non-empty (Boyd and Vandenberghe, 2004, §2.1.3) and Slater's condition holds (Boyd and Vandenberghe, 2004, §5.2.3). Consequently, there is no duality gap and the Karush-Kuhn-Tucker conditions are both necessary and sufficient for optimality (Boyd and Vandenberghe, 2004, §5.5.3). As the objective function of (A.5) is a quadratic form, it is a smooth function of the arguments $h$, $\lambda$, and $\Sigma(\varrho^*)$. Thus, the Karush-Kuhn-Tucker conditions stipulate that an optimal $\lambda$ is the root of continuous functions of $h$ and $\Sigma(\varrho^*)$. Since the solutions to the Karush-Kuhn-Tucker conditions are continuous functions of $h$ and $\Sigma(\varrho^*)$, the optima of (A.5) are continuous functions of $h$ and $\Sigma(\varrho^*)$.

Shapiro (2003) uses that strong duality holds for (A.4) to give rise to the identity

$$D^2_{\Lambda_+} = \sup_{\lambda \in \Lambda_+} \frac{\left[ \lambda^T \{ T - \mu(\varrho^*) \} \right]^2}{\lambda^T \Sigma(\varrho^*) \lambda}. \tag{A.6}$$

From this, it is seen by the definition of $A^2_{\Lambda_+}$ in (7) of the main text that $D^2_{\Lambda_+} = A^2_{\Lambda_+}$.

Shapiro (2003) shows that if $Y$ has a multivariate normal distribution with mean vector $\theta$ and known non-singular covariance matrix $V$ then

$$\sup_{\lambda \in \Lambda_+} \frac{\{ \lambda^T (Y - \theta) \}^2}{\lambda^T V \lambda} \sim \bar{\chi}^2(V^{-1}, \Lambda_+). \tag{A.7}$$

Since $I^{-1} \Sigma(\varrho^*) \to M$ as $I \to \infty$, it follows that $I^{1/2} \Sigma(\varrho^*)^{-1/2} \to M^{-1/2}$. By Lemma 2 the random vector $\Sigma(\varrho^*)^{-1/2} \{ T - \mu(\varrho^*) \}$ converges in distribution to a $K$-variate vector of independent standard normals. By Slutsky's

Lemma $I^{1/2}h$ converges in distribution to the mean-zero multivariate normal distribution with covariance $M^{-1}$. By continuity of the function taking $h$ to the optima of (A.5) along with (A.6), the mapping

$$\Sigma(\varrho^*)^{-1/2}\left\{T - \mu(\varrho^*)\right\} \mapsto \sup_{\lambda \in \Lambda_+} \frac{\left\{\lambda^T(T - \mu(\varrho^*))\right\}^2}{\lambda^T \Sigma(\varrho^*)\lambda}$$

is continuous. Exploiting Slutsky's Lemma, the Continuous Mapping Theorem, and (A.7) yields that $A^2_{\Lambda_+}$ converges in distribution to a $\bar{\chi}^2(M^{-1}, \Lambda_+)$ random variable as desired. $\square$

### A.1.4 Theorem 2

**Theorem 4.** *Suppose $\Lambda_1 \subseteq \Lambda_2$. Under mild conditions, the design sensitivity of (5) using $\Lambda = \Lambda_1$ is less than or equal to the design sensitivity of (5) using $\Lambda = \Lambda_2$.*

*Proof.* Define $\tilde{\Gamma}_\Lambda$ as the design sensitivity of the test using $a^*_{\Gamma,\Lambda}$ as a test statistic. To avoid triviality, suppose that the design sensitivities $\tilde{\Gamma}_{\Lambda_1}$ and $\tilde{\Gamma}_{\Lambda_2}$ both exist; see Rosenbaum (2004) and Rosenbaum (2013) for mild conditions for existence of the design sensitivity. Let $A_{\Gamma,\Lambda_i}$ be the random variable giving rise to the observation $a^*_{\Gamma,\Lambda_i}$ in (6) for $i = 1, 2$, that is

$$A^*_{\Gamma,\Lambda_i} = \min_{\varrho \in \mathcal{P}_\Gamma} \sup_{\lambda \in \Lambda_i} \frac{\lambda^T\{T - \mu(\varrho)\}}{\{\lambda^T \Sigma(\varrho)\lambda\}^{1/2}}.$$

Since $\Lambda_1 \subseteq \Lambda_2$, for any $\varrho$ we have

$$\sup_{\lambda \in \Lambda_1} \frac{\lambda^T\{T - \mu(\varrho)\}}{\{\lambda^T \Sigma(\varrho)\lambda\}^{1/2}} \leq \sup_{\lambda \in \Lambda_2} \frac{\lambda^T\{T - \mu(\varrho)\}}{\{\lambda^T \Sigma(\varrho)\lambda\}^{1/2}},$$

such that $A^*_{\Gamma,\Lambda_1} \leq A^*_{\Gamma,\Lambda_2}$. Consider any $\Gamma < \tilde{\Gamma}_{\Lambda_1}$. By the definition of design sensitivity, for a sensitivity analysis conducted at $\Gamma$ we have that $\text{pr}(A_{\Gamma,\Lambda_1} \geq k \mid \mathcal{Z})$ tends to one as $I \to \infty$ for any scalar $k$. Since $A^*_{\Gamma,\Lambda_1} \leq A^*_{\Gamma,\Lambda_2}$, for any $\Gamma < \tilde{\Gamma}_{\Lambda_1}$ the power of the test based upon $A^*_{\Gamma,\Lambda_2}$, $\text{pr}(A_{\Gamma,\Lambda_2} \geq k \mid \mathcal{Z})$, also tends to one as $I \to \infty$ for any $k$. Thus, $\tilde{\Gamma}_{\Lambda_2} \geq \tilde{\Gamma}_{\Lambda_1}$ as desired.

$\square$

## A.2 Additional Simulations

### A.2.1 The general setup of the simulation studies

In this section we present additional simulation studies to further illustrate the results presented in the manuscript. All of the simulation studies are conducted with some number $I$ pairs, and some number $K$ outcome variables, equicorrelated with correlation controlled by a parameter $\rho$. For each outcome variable, the employed test statistic is $T_k = \sum_{i=1}^I \text{sign}(Y_{ik}) \min(|Y_{ik}|/s_k, 2.5)$, where $s_k$ is the median of $|Y_{ik}|$ $(i = 1, .., I)$. This amounts to a choice of a $m$-statistic with Huber's $\psi$-function, as described in Rosenbaum (2007).

### A.2.2 Rejecting the global null with $I = 1000$ pairs

In § 5.2 of the main text the $\bar{\chi}^2$-test was compared to the equal-weight test and the test of Fogarty and Small (2016) with $I = 300$ matched pairs. To highlight the large-sample properties of the test, we include Figure A.1. As $I$ increases, the power curves converge pointwise to step functions, evaluating to 1 if $\Gamma$ is below the design sensitivity and zero otherwise (Rosenbaum, 2013). This indicates that the gap between the equal-weight test and the $\bar{\chi}^2$-test observed in the first row of Figure 1 and Figure A.1 will shrink as $I$ increases, and will disappear in the limit. This trend can be appraised visually by comparing the disparity observed in the first row of Figure 1 in the manuscript where $I = 300$ to the first row of Figure A.1 where $I = 1000$. As a consequence Theorem 2 and of the pointwise convergence of the power curve to the indicator function of the event $\Gamma$ less than the design sensitivity, the power curve of the $\bar{\chi}^2$-test will converge to that of the most powerful test at any fixed $\Gamma$ among all coherent tests.
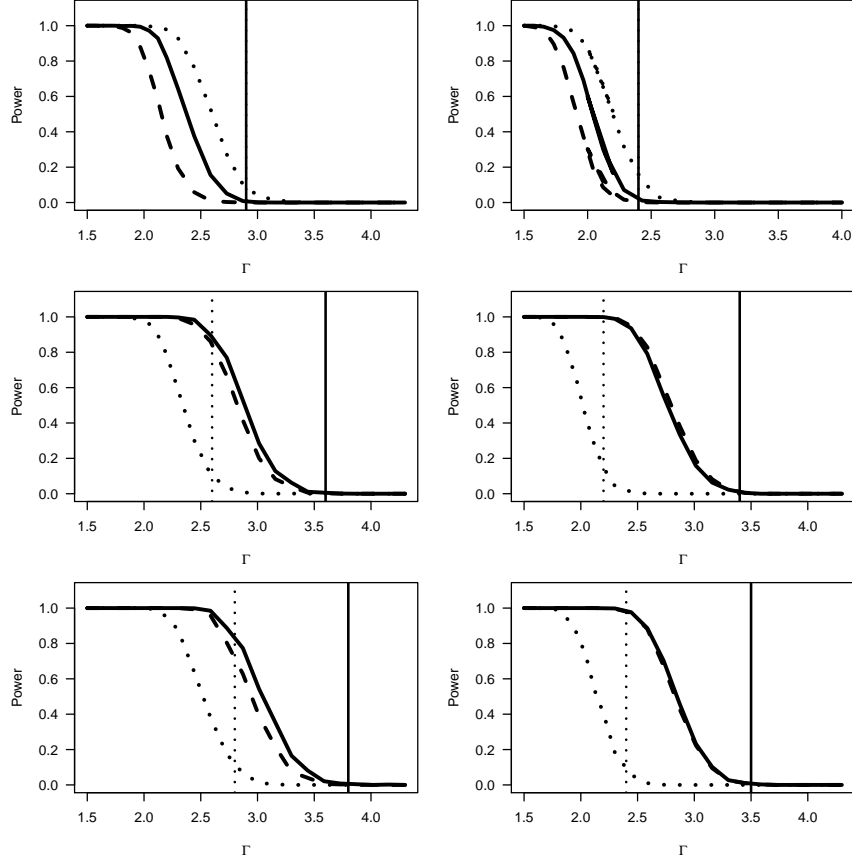
Figure A.1: Power comparisons between the method of Fogarty and Small (2016) (dashed), the method of this paper (solid), and the equal-weight test (dotted) as $\Gamma$ increases with $I = 1000$. The first row has $\tau_1 = \tau_2 = \tau_3 = 0.25$. The second row has $\tau_1 = \tau_2 = 0.1$ and $\tau_3 = 0.5$. The third row has $\tau_1 = 0.05$, $\tau_2 = 0.2$, and $\tau_3 = 0.5$. Figures on the left have $\rho = 0$ while on the right $\rho = 0.2$. For each fixed set of parameters, power simulations were performed on 1000 simulated data sets. The design sensitivity of the equal-weight test is the dotted vertical line and the design sensitivity of the $\bar{\chi}^2$-test is the solid vertical line. In the first row, these two design sensitivities are the same and are shown by the single solid vertical line.

### A.2.3 Rejecting individual nulls through closed testing

An experimenter may want to test not only the global null hypothesis $H_0$ of (4) but also the $K$ individual null hypotheses $H_1, \ldots, H_K$. To achieve this at level $\alpha$, she may use a closed-testing framework (Marcus et al., 1976). Then, in order to test $H_i$ at level $\alpha$, she performs $\alpha$-level tests all hypotheses of the form $H_i \wedge \left( \bigwedge_{k \in S_i} \right)$ with $S_i$ the set of all possible subsets of the numbers $1, \ldots, K$ excluding $i$; she then rejects $H_i$ if all of these tests rejected. Another standard method to test both the global null and each individual null would be to conduct a Bonferroni-corrected test of the global null and then use the results of the corrected individual tests to reject each $H_k$. Figure A.2 examines the performance of these two methods against the test of only $H_1$ when $\tau_1 = 0.5$, $\tau_2 = 0.2$, $\tau_3 = 0.05$ and equicorrelation between the paired differences at at $\rho = 0.2$. The comparison to the test of only $H_1$ is an unfair comparison in that testing only $H_1$ at level-$\alpha$ does not control the family-wise error rate at $\alpha$ when examining all $k = 1, \ldots, K$. However, the test of $H_1$ alone at level-$\alpha$ achieves the highest power possible for any testing procedure that tests $H_k$ as it does not employ any corrections to control the family-wise error rate. Thus, comparison to the test of $H_1$ alone at level-$\alpha$ serves as a comparison to an idealized benchmark, the absolute limit of statistical power that one may achieve when testing $H_1$ using a particular test statistic.

At all values of $I$ examined, the closed test outperforms the Bonferroni-corrected test. Furthermore, as $I$ increases, the closed test approaches the same power as the individual test without correction. Thus, for sufficiently large studies, empirical results suggest that a closed testing framework allows the experimenter to test both the global null and the
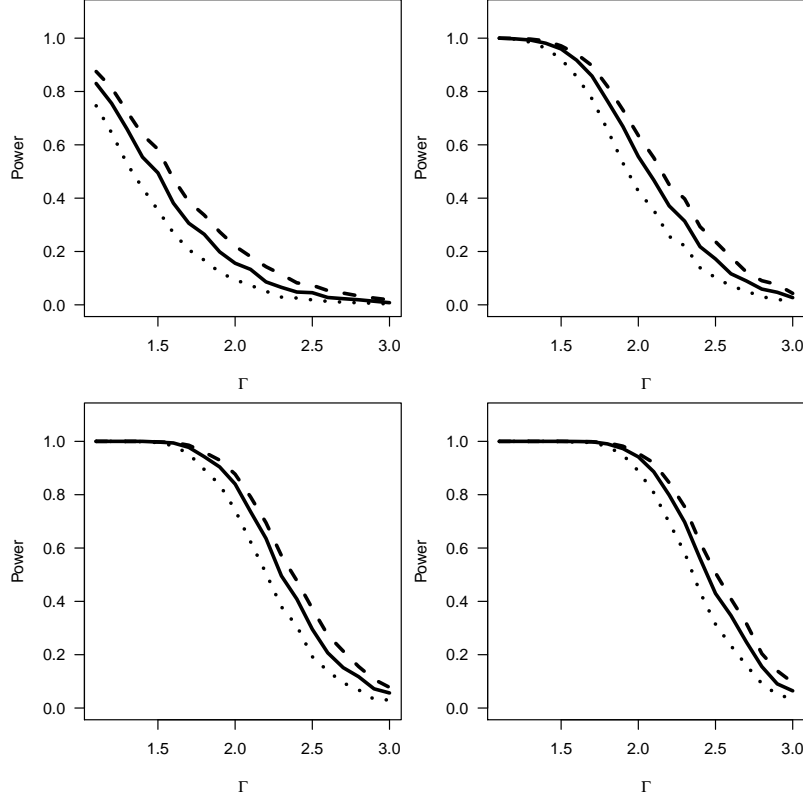
Figure A.2: Power comparisons between embedding the $\bar{\chi}^2$-test into a closed testing framework (solid), performing a Bonferroni-corrected test (dotted), and performing an uncorrected test (dashed) as $I$ increases. All simulations performed with $\tau_1 = 0.5$, $\tau_2 = 0.2$, $\tau_3 = 0.05$ with equicorrelation at $\rho = 0.2$ testing $H_1$. All data is with normal noise and tested with Huber's $\psi$-function as the underlying statistic. Additional parameters listed clockwise from the top-left: $I = 50$, $I = 150$, $I = 250$, and $I = 350$. For each sample size, power simulations were performed on 2000 simulated data sets.

| $\tau_2$ | $\rho$ | Fogarty and Small (2016) | $\Lambda = \Lambda_+$ | $\Lambda = \mathbb{R}^K \setminus \{0_K\}$ |
|---|---|---|---|---|
| -0.5 | 0 | 0 | 0 | 0.694 |
| -0.25 | 0 | 0 | 0 | 0.454 |
| 0 | 0 | 0.026 | 0.018 | 0.46 |
| -0.5 | 0.5 | 0 | 0 | 0.546 |
| -0.25 | 0.5 | 0 | 0 | 0.424 |
| 0 | 0.5 | 0.018 | 0.016 | 0.496 |

Table A.1: Type I error rates for the method of Fogarty and Small (2016), the $\bar{\chi}^2$-test of this paper, and the test taking $\Lambda = \mathbb{R}^K \setminus \{0_K\}$ using $\alpha = 0.05$. All tests performed with $I = 20$ matched pairs, $\tau_1 = -0.5$, and $\Gamma = 1$. For each set of parameters the power was estimated based upon 500 simulations.

individual null at level-$\alpha$ with minimal loss of power from multiple comparisons relative to testing only the individual null.

### A.2.4  Type I error control in small samples using the asymptotic reference distribution

In this simulation, we assess the Type I error rate with $I = 20$ matched pairs at $\Gamma = 1$. In each simulation, the global null of non-positive treatment effects is true. Table A.1 details simulated Type I error rates for the method of Fogarty and Small (2016), the $\bar{\chi}^2$-test of this paper, and the unconstrained test taking $\Lambda = \mathbb{R}^K \setminus \{0_K\}$ for $K = 2$ outcome variables with a range of different parameter values.

Both the method of Fogarty and Small (2016) and the $\bar{\chi}^2$-test control the Type I error rate at $\alpha$ even when in the finite sample regime while using the asymptotic reference distribution. As alluded to in §4.1 the test taking $\Lambda = \mathbb{R}^K \setminus \{0_K\}$ fails to control the Type I error rate at $\alpha$ since allowing $\lambda$ to have an unconstrained sign in each coordinate removes the ability to discriminate positive treatment effects from negative treatment effects. This further motivates the restriction to the set of coherent combinations $\Lambda_+$.

## A.3  Algorithmic Details for Conducting the Sensitivity Analysis

The optimization problem in (6) is solved via a projected subgradient descent algorithm. Shor (1985) contains a detailed introduction to subgradient methods. The algorithm begins with some initial feasible $\varrho_{(0)}$, solves for an optimal $\lambda$ under the fixed $\varrho_{(0)}$, computes a subgradient of the objective at the optimal $\lambda$, and uses the subgradient to project onto the feasible region thereby locating a $\varrho_{(1)}$. The procedure iterates until convergence criteria are satisfied.

Formally, given a feasible $\rho_{(n)}$ we compute

$$\lambda^*_{\rho_{(n)}} = \sup_{\lambda \in \Lambda_+} \frac{\left[\max\{0, \lambda^T(T - \mu(\rho_{(n)}))\}\right]^2}{\lambda^T \Sigma(\rho_{(n)})\lambda}. \tag{A.8}$$

To compute (A.8), results from Shapiro (2003) are leveraged to allow efficient computation of

$$\sup_{\lambda \in \Lambda_+} \frac{\lambda^T\{T - \mu(\rho_{(n)})\}}{\{\lambda^T \Sigma(\rho_{(n)})\lambda\}^{1/2}},$$

by solving a single quadratic program. In the event that

$$\sup_{\lambda \in \Lambda_+} \frac{\lambda^T\{T - \mu(\rho_{(n)})\}}{\{\lambda^T \Sigma(\rho_{(n)})\lambda\}^{1/2}} > 0$$

$\lambda^*_{\rho_{(n)}}$ is set to the optimizing choice of $\lambda$. However, when

$$\sup_{\lambda \in \Lambda_+} \frac{\lambda^T\{T - \mu(\rho_{(n)})\}}{\{\lambda^T \Sigma(\rho_{(n)})\lambda\}^{1/2}} \leq 0$$

there exists a feasible $\varrho$ such that the test fails to reject the sharp null, and thus no further iterations of the subgradient method are needed.

By Hiriart-Urruty and Lemarechal (2013), if $f(x) = \sup_{j \in J} f_j(x)$ where each $f_j(x)$ is a convex function, $f(x) = f_{j^*}(x)$, and $g \in \partial f_{j^*}(x)$, then $g \in \partial f(x)$. In less technical terms, to compute a subgradient of a function which is the point-wise supremum of a many convex functions, one first finds a function $f_{j^*}(\cdot)$ which achieves the maximum value at the desired point $x$ and then one computes a subgradient of this function. As such, at the optimal value $\lambda^*$ one computes that the subgradient of the objective function with respect to the variables $\varrho_i = (\varrho_{i1}, \ldots, \varrho_{in_i})$ is

$$g = \frac{h_1(\varrho)\partial_{\varrho_i} h_2(\varrho) - h_2(\varrho)\partial_{\varrho_i} h_1(\varrho)}{h_2(\varrho)^2},$$

where

$$h_1(\varrho) = \left(\lambda^{*T}(T - \mu(\varrho))\right)^2,$$
$$h_2(\varrho) = \lambda^{*T} \Sigma(\varrho)\lambda^*,$$
$$\partial_{\varrho_i} h_1(\varrho) = -2(Q_i^T \lambda^*)\lambda^{*T}(T - \mu(\varrho)),$$
$$\partial_{\varrho_i} h_2(\varrho) = (Q_i^T \lambda^*) \circ (Q_i^T \lambda^*) - 2(Q_i^T \lambda^*)(Q_i^T \lambda^*)^T \varrho_i,$$

where $Q_i$ is the $K$-by-$n_i$ matrix where the $(k, j)$th entry is $q_{ijk}$ and $\circ$ denotes the coordinate-wise product operation.

Armed with the solution to the inner maximization and the form of the subgradient $g$, we can now detail the projected subgradient descent method.

1. Initialize a feasible $\rho_{(0)}$, pick $t_0 > 0$ and $n = 1$
2. Repeat until convergence:
   (a) Find $\lambda_{\rho_{(n-1)}}$ by solving (A.8),

    (b) Compute the subgradient $g$ from (A.3) using $\lambda_{\rho_{(n-1)}}$,

    (c) Define $\rho_{(n)}$ to be the projection of $\rho_{(n-1)} - t_{n-1}g$ onto the feasible region,

    (d) Update the parameters: $t_n = t_0/\sqrt{n}$ and $n = n + 1$.

Since the objective function is convex and the feasible set is also convex, any local optimum is a global optimum as well. In practical execution on both synthetic and real data sets convergence has been observed after few iterations.

## A.4 The $\bar{\chi}^2$ Distribution

### A.4.1 Finding a better critical value

While the subgradient method solves (6) and Theorem 1 gives that the asymptotic distribution of $A^2_{\Lambda_+}$ is $\bar{\chi}^2$, the weights of the limiting distribution are still unknown. Comparing the value of (6) against the $1 - \alpha$ quantile arising from the bound

$$\mathrm{pr}\{\bar{\chi}^2(V, \Lambda_+) \geq c\} \leq 0.5\{\mathrm{pr}\left(\chi^2_{K-1} \geq c\right) + \mathrm{pr}\left(\chi^2_K \geq c\right)\} \tag{A.9}$$

would control the Type I error. While improving over a critical value based on a $\chi^2_K$ distribution, the bounds through (A.9) are still unduly conservative. We now describe an algorithm which exploits the particular structure of the sensitivity analysis problem to dramatically improve the critical value.

By directly computing upper and lower bounds on the correlation between $T_k$ and $T_\ell$ for each $k, \ell = 1, \ldots, K$ one can compute coordinate-wise upper and lower bounds on on the overall correlation matrix $\mathrm{diag}\{\Sigma(\varrho)\}^{-1/2}\Sigma(\varrho)\mathrm{diag}\{\Sigma(\varrho)\}^{-1/2}$, where $\mathrm{diag}\{\Sigma(\varrho)\}$ contains the diagonal elements of $\Sigma(\varrho)$ on its diagonals but has zeroes on its off-diagonals. Since the weights of the $\bar{\chi}^2$ distribution depend on $\Sigma(\varrho)$ only through its correlation matrix (Silvapulle and Sen, 2005), one can directly optimize over bounds on the marginal correlations to find the most conservative $1 - \alpha$ critical value associated to a correlation matrix within the bounds. This optimization can be performed via either numerical approximation of gradients or by directly computing gradients of the $p$-value function with respect to the correlations. Such gradients are accessible due to Plackett's identity (Plackett, 1954) and can be calculated with assistance of functions in the `mvtnorm` package within R for evaluating orthant probabilities and the density of the multivariate normal. Optimizing over the space of correlation matrices yields significant improvement over the critical value drawn from previous bound. Figure A.3 highlights the differences between using the $1 - \alpha$ quantile from a $\chi^2$ distribution, the $1 - \alpha$ quantile from the naive bound based upon (A.9), and using the optimal $1 - \alpha$ quantile with $K = 3$ outcome variables. By using the most conservative $1 - \alpha$ quantile within the upper and lower bounds on the correlation matrix the Type I error rate is asymptotically controlled at $\alpha$.

There is a true, but generally unknown, underlying correlation structure between the test statistics that depends upon the true vector of conditional probabilities $\varrho^*$. Thus, the true $1 - \alpha$ quantile from the $\bar{\chi}^2$ distribution with weights based on the true correlation would not change with the value of $\Gamma$ employed in the sensitivity analysis. As the true unmeasured confounders are unknown, we instead find a conservative critical value based upon the feasible values for $\varrho$ at a given $\Gamma$. As $\Gamma$ grows so too does the feasible region for the probabilities $\mathcal{P}_\Gamma$; consequently the conservative critical value increases with $\Gamma$ as well. This explains the trend in the right-hand panel of Figure A.3.

### A.4.2 The worst-case correlation with bivariate outcomes

In the case for $K = 2$, an elementary proof establishes a closed form of the optimizing correlation matrix subject to box constraints.

**Theorem 5.** *Suppose that $K = 2$ and $[\ell, u] \subseteq (-1, 1)$. Over all matrices $M$ in the set*

$$S = \left\{ \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} : \rho \in [\ell, u] \right\}$$

*the matrix $\begin{bmatrix} 1 & \ell \\ \ell & 1 \end{bmatrix}$ achieves the most conservative (largest possible) $1 - \alpha$ quantile of $\bar{\chi}^2(M^{-1}, \Lambda_+)$.*

*Proof.* Say that $X \sim \bar{\chi}^2(M^{-1}, \Lambda_+)$ for $M^{-1} \in S$. From Sen and Silvapulle (2002) the probability

$$\mathrm{pr}\left(X \leq c\right) = w_0(2, M^{-1})\mathrm{pr}\left(\chi^2_0 \leq c\right) + \frac{1}{2}\mathrm{pr}\left(\chi^2_1 \leq c\right) + w_2(2, M^{-1})\mathrm{pr}\left(\chi^2_2 \leq c\right)$$

$$= w_2(2, M)\mathrm{pr}\left(\chi^2_0 \leq c\right) + \frac{1}{2}\mathrm{pr}\left(\chi^2_1 \leq c\right) + w_0(2, M)\mathrm{pr}\left(\chi^2_2 \leq c\right)$$
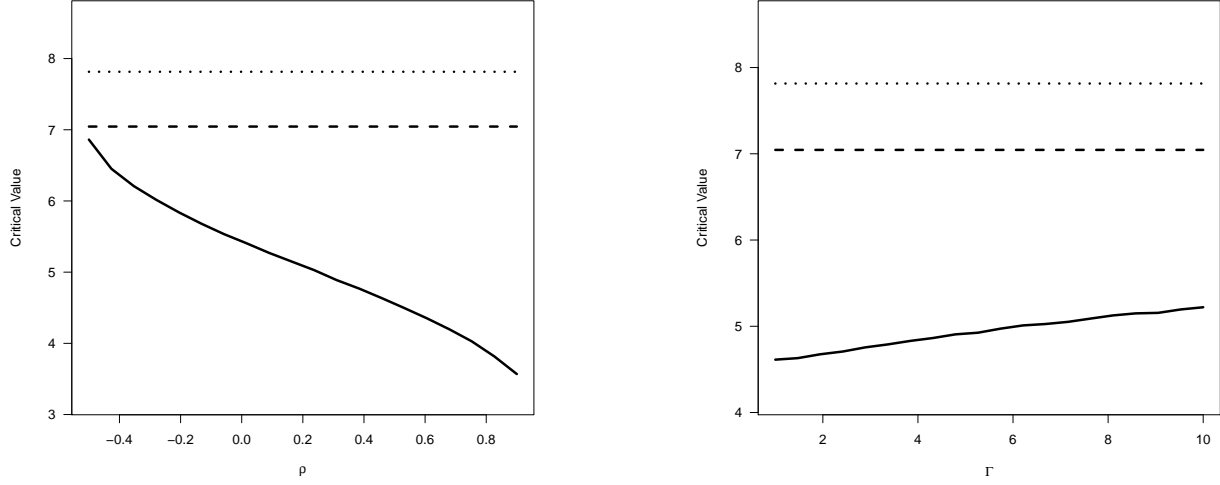
Figure A.3: $1 - \alpha$ quantiles for $\alpha = 0.05$ generated for the trivariate scenario $I = 300$, $\tau_1 = \tau_2 = \tau_3 = 0$. On the left, $\Gamma$ is fixed at 1 while $\rho$ varies over $[-0.5, 0.9]$. On the right, $\rho$ is fixed at 0.5 while $\Gamma$ ranges from 1 to 10. In both figures the $\chi_3^2$ $1 - \alpha$ quantile is the dotted line, that of the naive bound derived from (A.9) is the dashed line, and the $1 - \alpha$ quantile coming from optimizing over feasible correlation matrices is the solid line.

where each $\chi_i^2$ is an independent random variable with $\chi_i^2$ distribution. The value $w_{2-i}(2, M)$ is the probability that the projection, under the norm induced by the quadratic form $x^T M x$, of a standard bivariate normal random vector onto the non-negative orthant has exactly $2 - i$ positive components. By an argument presented in Sen and Silvapulle (2002), this interpretation of $w_{2-i}(2, M)$ is equivalent to defining $w_{2-i}(2, M)$ as the probability that a standard bivariate normal random variable $Z$ falls into $R_i = \left\{ x \in \mathbb{R}^2 \mid \sum_{k=1}^2 1(b_k > 0) = i \right\}$ where $b = M^{1/2} z$. Since $w_2(2, M) + w_0(2, M) = 1$ and $\text{pr}\left(\chi_0^2 \leq c\right) \geq \text{pr}\left(\chi_2^2 \leq c\right)$ for all scalars $c$ it suffices to maximize $w_2(2, M)$. Taking

$$M = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

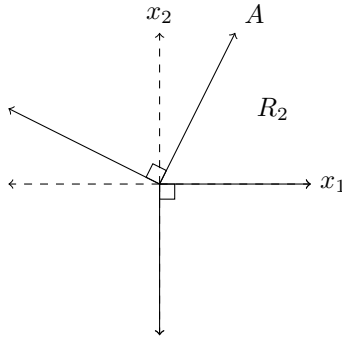gives that maximizing $w_2(2, M)$ is equivalent to maximizing the area $R_2$ in Figure A.4



Figure A.4: Pictorial representation of the region $R_2$. The upper right boundary of $R_2$ is the line given by $A$.

The slope of the line $A$ is $\rho - \rho^{-1}$ when $\rho \neq 0$ and $A$ is vertical when $\rho = 0$. Maximizing $R_2$ corresponds to taking $\rho$ as small as possible within $[\ell, u]$. Thus the matrix $\rho = \ell$ achieves the most conservative $1 - \alpha$ critical value of $\bar{\chi}^2(M^{-1}, \Lambda_+)$.

$\square$

### A.4.3 A bivariate illustration of the $\bar{\chi}^2$ distribution

Consider a mean-zero bivariate normal with covariance $V$ and consider the distribution of $\bar{\chi}^2(V, \Lambda_+)$. By the law of total probability, $\mathrm{pr}\left(\bar{\chi}^2(V, \Lambda_+) \leq c\right) = \sum_{i=0}^{2} \mathrm{pr}\{\bar{\chi}^2(V, \Lambda_+) \leq c \mid X \in R_i\}\mathrm{pr}(X \in R_i)$ where $R_0, R_1$, and $R_2$ are disjoint coverings of $\mathbb{R}^2$. Let $b = V^{-1/2}x$, and set $R_i = \left\{x \in \mathbb{R}^2 \mid \sum_{k=1}^{2} 1(b_k > 0) = i\right\}$; this is shown in Figure A.5.
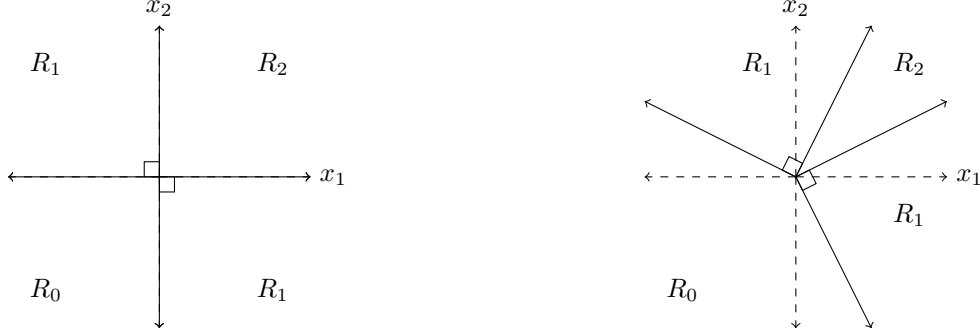
Figure A.5: The regions corresponding to different distributional forms of the likelihood ratio statistic. In the left image $V = I_{2 \times 2}$; the right image illustrates the general case, in this case the correlation is $-0.8$.

Within each $R_i$, $\bar{\chi}^2(V, \Lambda_+) \sim \chi_i^2$, where $\chi_0^2$ is a point mass at zero. The weights of the $\bar{\chi}_K^2(V, \Lambda_+)$ are determined by the probability of falling into each partition, and are seen to depend on the covariance $V$.

Expansive literature exists on the $\bar{\chi}^2$ distribution. The paper Kudô (1963) introduces the topic in the context of order constrained one-sided tests; Chapter 3 of Silvapulle and Sen (2005) contains detailed examples and derivations as well as a collection of many contemporary results; and Shapiro (1985) discusses the weights $w_i(k, V, C)$ extensively.

## A.5 Matching Details for Smoking and Polycyclic Aromatic Hydrocarbons

Individuals were classified as cigarette smokers or as non-cigarette-smokers in accordance with the criteria used in Fogarty and Small (2016). This divided the population of 1638 total individuals into 432 cigarette smokers and 1206 non-cigarette-smokers. The population of non-smokers did include those who may have smoked in the past but had stopped smoking by the time of the survey, as well as individuals who had never smoked cigarettes. The individuals were placed into matched groups using a full matching procedure (Rosenbaum, 2010a, §8.5); thus each group contained a single treated unit and multiple control units or a single control unit and multiple treated units. Pre-treatment covariates were selected based upon recent medical research. To form the fully-matched sets, propensity score caliper with a rank-based Mahalanobis distance for within-caliper distance was used. The caliper was set at 0.08 and logistic regression was performed to estimate propensity scores (Rosenbaum, 2010a, §8). See Fogarty and Small (2016, Appendix A) for further implementation details.

## References

Carl-Elis Boström, Per Gerde, Annika Hanberg, Bengt Jernström, Christer Johansson, Titus Kyrklund, Agneta Rannug, Margareta Törnqvist, Katarina Victorin, and Roger Westerholm. Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. *Environmental Health Perspectives*, 110:451–488, 2002.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004. ISBN 0-521-83378-7. URL https://doi.org/10.1017/CBO9780511804441.

A. Charnes and W. W. Cooper. Programming with linear fractional functionals. *Naval Res. Logist. Quart.*, 9:181–186, 1962. ISSN 0028-1441. doi: 10.1002/nav.3800090303. URL https://doi.org/10.1002/nav.3800090303.

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

Colin B Fogarty. On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:1035–1056, 2018.

Colin B. Fogarty and Dylan S. Small. Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *J. Amer. Statist. Assoc.*, 111(516):1820–1830, 2016. ISSN 0162-1459. URL `https://doi.org/10.1080/01621459.2015.1120675`.

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

Ben B Hansen. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618, 2004. doi: 10.1198/016214504000000647. URL `https://doi.org/10.1198/016214504000000647`.

Jean-Baptiste Hiriart-Urruty and Claude Lemarechal. *Convex analysis and minimization algorithms I: Fundamentals.*, volume 305. Springer Science and Business Media, 2013.

Akio Kudô. A multivariate analogue of the one-sided test. *Biometrika*, 50:403–418, 1963. ISSN 0006-3444. URL `https://doi.org/10.1093/biomet/50.3-4.403`.

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

Michael D Perlman. One-sided testing problems in multivariate analysis. *The Annals of Mathematical Statistics*, 40(2):549–567, 1969.

R. L. Plackett. A reduction formula for normal multivariate integrals. *Biometrika*, 41:351–360, 1954. ISSN 0006-3444. doi: 10.1093/biomet/41.3-4.351. URL `https://doi.org/10.1093/biomet/41.3-4.351`.

C. Radhakrishna Rao. *Linear statistical inference and its applications*. New York: John Wiley & Sons, second edition, 1973.

Tim Robertson, FT Wright, and RL Dykstra. *Order restricted statistical inference.* John Wiley & Sons, New York, 1988.

Paul R Rosenbaum. Quantiles in nonrandom samples and observational studies. *Journal of the American Statistical Association*, 90(432):1424–1431, 1995.

Paul R Rosenbaum. Signed rank statistics for coherent predictions. *Biometrics*, 53(2):556–566, 1997.

Paul R Rosenbaum. *Observational studies*. Springer, New York, 2002.

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

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

Paul R. Rosenbaum. *Design of observational studies*. Springer Series in Statistics. Springer, New York, 2010a. ISBN 978-1-4419-1212-1. URL `https://doi.org/10.1007/978-1-4419-1213-8`.

Paul R Rosenbaum. *Design of observational studies*. Springer, New York, 2010b.

Paul R. Rosenbaum. Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics*, 69(1):118–127, 2013. ISSN 0006-341X. doi: 10.1111/j.1541-0420.2012.01821.x. URL `https://doi.org/10.1111/j.1541-0420.2012.01821.x`.

Paul R Rosenbaum. How to see more in observational studies: Some new quasi-experimental devices. *Annual Review of Statistics and Its Application*, 2:21–48, 2015.

Paul R Rosenbaum. Using Scheffé projections for multiple outcomes in an observational study of smoking and periodontal disease. *The Annals of Applied Statistics*, 10(3):1447–1471, 2016.

Paul R. Rosenbaum. Sensitivity analysis for stratified comparisons in an observational study of the effect of smoking on homocysteine levels. *Annals of Applied Statistics*, to appear, 2018.

Henry Scheffé. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2):87–110, 1953.

Pranab K. Sen and Mervyn J. Silvapulle. An appraisal of some aspects of statistical inference under inequality constraints. *J. Statist. Plann. Inference*, 107(1-2):3–43, 2002. ISSN 0378-3758. URL `https://doi.org/10.1016/S0378-3758(02)00242-2`.

WR Shadish, Thomas D Cook, and DT Campbell. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Wadsworth, Belmont, 2002.

Alexander Shapiro. Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika*, 72(1):133–144, 1985. ISSN 0006-3444. doi: 10.1093/biomet/72.1.133. URL `https://doi.org/10.1093/biomet/72.1.133`.

Alexander Shapiro. Scheffe's method for constructing simultaneous confidence intervals subject to cone constraints. *Statist. Probab. Lett.*, 64(4):403–406, 2003. ISSN 0167-7152. URL `https://doi.org/10.1016/S0167-7152(03)00205-0`.

N. Z. Shor. *Minimization methods for nondifferentiable functions*, volume 3 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1985. ISBN 3-540-12763-1. doi: 10.1007/978-3-642-82118-9. URL `https://doi.org/10.1007/978-3-642-82118-9`. Translated from the Russian by K. C. Kiwiel and A. Ruszczyński.

Mervyn J. Silvapulle and Pranab K. Sen. *Constrained statistical inference*. Hoboken: John Wiley & Sons, 2005. ISBN 0-471-20827-2.

Scott L. Tomar and Samira Asma. Smoking-attributable periodontitis in the United States: Findings from NHANES III. *Journal of Periodontology*, 71(5):743–751, 2000. doi: 10.1902/jop.2000.71.5.743. URL `https://onlinelibrary.wiley.com/doi/abs/10.1902/jop.2000.71.5.743`.

José R Zubizarreta. Using mixed integer programming for matching in an observational study of kidney failure after surgery. *Journal of the American Statistical Association*, 107(500):1360–1371, 2012.