ASSESSMENT OF THE EXTENT OF CORROBORATION OF AN ELABORATE THEORY OF A CAUSAL HYPOTHESIS USING PARTIAL CONJUNCTIONS OF EVIDENCE FACTORS

By Bikram Karmakar and Dylan S. Small

University of Pennsylvania, Philadelphia

An elaborate theory of predictions of a causal hypothesis consists of several falsifiable statements derived from the causal hypothesis. Statistical tests for the various pieces of the elaborate theory help to clarify how much the causal hypothesis is corroborated. In practice, the degree of corroboration of the causal hypothesis has been assessed by a verbal description of which of the several tests provides evidence for which of the several predictions. This verbal approach can miss quantitative patterns. In this paper, we develop a quantitative approach. We first decompose these various tests of the predictions into independent factors with different sources of potential biases. Support for the causal hypothesis is enhanced when many of these evidence factors support the predictions. A sensitivity analysis is used to assess the potential bias that could make the finding of the tests spurious. Along with this multi-parameter sensitivity analysis, we consider the partial conjunctions of the tests. These partial conjunctions quantify the evidence supporting various fractions of the collection of predictions. A partial conjunction test involves combining tests of the components in the partial conjunction. We find the asymptotically optimal combination of tests in the context of a sensitivity analysis. Our analysis of an elaborate theory of a causal hypothesis controls for the familywise error rate.

MSC 2010 subject classifications: Primary 62G10; secondary 62K15 03A10

Keywords and phrases: Causal inference, degree of corroboration, elaborate theory, evidence factors, observational studies

1. Introduction

1.1. An elaborate theory of a causal effect and evidence factors

Fisher's response to the question "what can be done in observational studies to clarify the step from association to causation[?]" was: "Make your theories elaborate"(Cochran, 1965). Cochran explains this response by stating that to clarify the step from association to causation one should envision as many different consequences as possible of the causal hypothesis under investigation and design studies which are able to scrutinize these consequences. In parallel to Cochran's interpretation of Fisher's response, Popper (1934, 1972), through arguments of classical logic, emphasizes the importance to scientific progress for a hypothesis to have a higher 'degree of testability'. By degree of testability, Popper means the amount of falsifiable 'basic statements' the theory generates. "If we look for confirmations", Popper (1963) writes, "It is easy to obtain confirmations ... for nearly every theory", while "[e]very genuine test of a theory is an attempt to falsify it, or to refute it. Testability is falsifiability[.]"

The motivating example of this paper, discussed in detail in §2, considers the causal hypothesis that exposure to lead of a parent at the workplace causes high level of lead in the blood of a child at home. To test this causal hypothesis, Morton et al. (1982) established the following elaborate theory (Rosenbaum, 2005): (a) children of parents who were occupationally exposed to lead will have higher lead levels in the blood than otherwise similar control children; (b) among children of parents occupationally exposed to lead, children of parents with higher occupational lead exposure will have higher lead levels than otherwise similar children of parents with lower occupational lead exposure; and (c) among children of parents occupationally exposed to lead, children whose parents practiced poorer hygiene before leaving work will have higher lead levels than otherwise similar children whose parents practiced better hygiene. We are interested in the question: what is the extent of corroboration of this theory provided by the data? Popper (1972), in the addendum to his final chapter of *The Logic of Scientific Discovery*, writes, "I tried to make clear that by the *degree of corroboration* of a theory I mean a brief report that summarizes the way in which the theory has stood up to tests, and how severe these tests were." In practice, the degree of corroboration of an elaborate theory has been evaluated by reporting what fraction of test of predictions of the elaborate theory have p-values < 0.05 (where rejecting the null supports the elaborate theory); see, e.g., Centerwall (1989) or Wong, Cook and Steiner (2015).

There are two problems with just counting the fraction of p-values less than 0.05 for assessing degree of corroboration of an elaborate theory. First, if the tests are dependent, then multiple tests rejecting may not be providing much more evidence than one test rejecting. Second, counting the fraction of p-values less than 0.05 is not an efficient combination of the evidence. For example, if two independent tests of the same null hypothesis both have p-values 0.06, this is strong evidence against the null by Fisher's method of combining independent tests (Fisher, 1932), the p-value for Fisher's combined test is 0.02.

An additional problem with the current practice for assessing the degree of corroboration for an elaborate theory is that the p-value computed for each test of the elaborate theory assumes no unmeasured confounding. In most observational studies, unmeasured confounding is a concern, and we would not find convincing an inference that was valid with no unmeasured confounding but invalid with a little bit of unmeasured confounding. A sensitivity analysis examines how much bias from unmeasured confounding could change the conclusions of a study that assumed no unmeasured confounding (Cornfield et al., 1959; Rosenbaum, 1987; Hosman et al., 2010; Keele and Minozzi, 2013; Stuart et al., 2013; Ding and Vanderweele, 2016; Fogarty and Hasegawa, 2018).

We develop a method for assessing the extent of corroboration of an elaborate theory that overcomes the three shortcomings we identified above of the current p-value counting approach. Our method involves three aspects: (i) we decompose the test of the elaborate theory into evidence factors, pieces that are affected by different biases and statistically near independent (Rosenbaum, 2011, 2017; Zubizarreta et al., 2012) (the additional requirement of different biases in each test increases robustness of the analysis against multiple potential sources of biases); (ii) we assess the extent of corroboration in a way that combines the information from different tests efficiently and furthermore we use partial conjunction tests (Benjamini and Heller, 2008; Benjamini, Heller, and Yekutieli, 2009); and (iii) we test the evidence factors using sensitivity analysis methods that allow for specified amounts of unmeasured confounding. The novel contributions of the paper are the following: (a) we provide a systematic approach to decomposing an elaborate theory into evidence factors; (b) as a way to test for partial corroboration of the elaborate theory, we introduce partial conjunction tests (partial conjunction tests have been previously developed for the purpose of inference in neuroimaging experiments by Benjamini and Heller, 2008); (c) we develop a sensitivity analysis method for carrying out (a) and (b) that allows for a specified degree of unmeasured confounding; (d) we show

that the method developed for (c) controls for the overall familywise error rate in the multi-parameter sensitivity analysis; and (e) for the method for (c), which involves combining sensitivity analyses for each of the evidence factors, we find the asymptotically optimal such combining method.

1.2. Sir Karl Popper and degree of corroboration

The term 'degree of corroboration' was introduced by Popper in response to an inattentive translation, 'degree of confirmation', of his original phrase '*Grad der Bewährung*'. Two decades after *Logik der Forschung*, in three *Br. J. Philos. Sci.* notes (vol. **5**, pp. 143–149, 1954; vol. **7**, pp. 350–353, 1957; and vol. **8**, pp. 294–302, 1958) Popper came up with a definition of *degree of confirmation* or *degree of corroboration*. In these notes, his motivation was rather different. He first attempted to show that, in the sense it is to be used in science, *degree of corroboration or acceptability of a theory* cannot be a probability. After showing this, he suggested a definition of *the degree to which a statement x is confirmed by a statement y* which he named the *degree of confirmation of x by y*. This definition was based on a list of *desiderata* he had put down for such a quantity. This definition may serve its purpose, but does not serve ours. First, such a definition depends on a background probability measure appropriately defined on first-order languages, and computations under this probability measure have not been well developed for statistical practice (Popper, 1954; Crupi, Chater and Tentori, 2013). Second, it is still an unsettled debate whether such a quantity is an adequate measure of corroboration (Rowbottom, 2013; Sprenger, 2018). Finally, this definition attempts to answer a very different question than ours. We are interested in the investigation of a causal hypothesis in an observational study and how best to make inferences about it from a frequentist perspective, whereas Popper attempted to define a quantity which would replace the p-value in investigation of a scientific theory.

1.3. Outline of the paper

The paper is organized as follows. We discuss our motivating example in §2. Here we briefly recall the original study. The notation for our method is introduced in §3.1. Section 3.2 recalls the treatment assignment models for the observed data. A brief review of the testing procedures and their sensitivity analysis is given in §3.3. The decomposition of the tests into evidence factors is established in §4. Our main method is developed in §5. In particular, Proposition 2 defines the (maximum) p-values for tests of partial conjunction of the hypotheses. Using these p-values we get tests of all the

partial conjunctions of the hypotheses for any given value of the sensitivity parameters. Theorem 3 and its corollaries show that the familywise error rate is controlled in our multi-parameter sensitivity analysis, with a range of values of the bias parameter, for the tests of the collection of all the partial conjunctions of the hypotheses. Section 6 compares the methods of testing the elaborate theory in their performance in sensitivity analysis. Section 6.2 finds asymptotically optimal methods in sensitivity analysis for tests of partial conjunctions of the hypotheses for elaborate theories. In §6.3, a simulation study is used for comparison of various methods in their power of sensitivity analysis. The simulation show that methods that pool evidence from the various evidence factors are favorable over methods that look at the individual tests who lose power when looking at fractions of the elaborate theory. Results of the study in §2 are in §7 and the paper ends with a short conclusion in §8.

2. Lead absorption study of Morton et al.

2.1. The elaborate theory and the analysis

Morton et al. (1982) studied the effect on children of a parent's occupational exposure to lead. Does exposure of a parent, who works in a battery manufacturing plant (in Oklahoma), to lead at the workplace cause an increase in lead level in the blood of a child in the household? The causal hypothesis is that an employee who is exposed to lead at the workplace carries lead dust back to the household and causes the child to have a higher lead level. To study their elaborate theory, given in §1.1, they collected data on 33 matched pairs, with one exposed child and one control child forming a pair. Data were collected on the lead level in the blood of the children; on the lead exposure levels, at the workplace, of the parents of the exposed children — categorized as high, medium, and low; and on hygiene practices of the parents of the exposed children before leaving work — categorized as good, moderately good, and poor.

A multitude of tests were carried out to see if the observed data are consistent with various pieces of the elaborate theory. They found a significantly higher lead level in exposed children compared to their controls. Exposed children of parents with higher lead exposure seemed to have higher lead levels, and parent's better hygiene practices seemed to indicate a lower lead level in the blood of the children. Focus was not on the separate pieces of the analyses but on the fact that there was a tendency of the evidence to converge to the same direction of confirming the elaborate theory. Although

not all the tests corroborated the elaborate theory, e.g., in comparing the exposed children depending on their parent's lead exposure level, 'the medium exposure group was not significantly different from the low exposure group,' the concluding remark of the authors was that the study 'provides additional confirmation that increased risk of lead absorption occurs in children of employees in a lead-related industry[.]' Clearly, the strategy was of a multiplist (Reynolds and West, 1987) — several pieces of evidence seeming to converge in favor of the causal hypothesis has been taken as a confirmation of the hypothesis. We will develop a more quantitative approach to summarizing the evidence about the elaborate theory from the study.

2.2. Is there evidence for a causal effect on children of occupational exposure to lead?

Wilcoxon's signed rank test for a higher lead level in the blood for exposed child compared to its control has a p-value $P_1 = 6.96 \cdot 10^{-5}$. Among the exposed children, the p-value in comparing high or moderate lead exposure at workplace for the parent versus a low exposure, using Wilcoxon's rank sum test, is $P_2 = 3.81 \cdot 10^{-3}$. A comparison of exposed children with high lead exposure level of the parents to medium lead exposure level of the parents is $P_3 = 9.59 \cdot 10^{-2}$. Of these three comparisons, the first one tests part (a) of the elaborate theory, the latter two are tests for part (b) of the elaborate theory. For part (c), consider exposed children from families with parent exposed to high level of lead. The p-value is $P_4 = 9.44 \cdot 10^{-3}$ when comparing poor hygiene practice versus a good or moderately good hygiene practice, and the p-value is $P_5 = 0.42$ in comparing a moderately good to a good hygiene practice. Note that for each test, a prediction of a true causal hypothesis is set up as an alternative hypothesis.

If we ask for evidence that all pieces of the elaborate theory are true, we would look at the maximum of those five p-values, which is 0.42. However, if were to pool all the p-values using Fisher's method — which will be shown using Theorem 1 gives a valid p-value — the pooled p-value is $1.41 \cdot 10^{-6}$, evidence in support of the hypothesis that at least one part of the elaborate theory is true. These are two drastically different numbers — neither suffices for our requirement of representing the extent of corroboration of the elaborate theory offered by the study. If we use the Holm-Bonferroni procedure, it would say that, at level 0.05, there is evidence to reject three out of the five tests, since $(5 + 1 - 3)P_{(3)} = 0.02832 < 0.05$ and $(5 + 1 - 4)P_{(4)} = 0.191846 > 0.05$ (Holm, 1979). We provide the results from our method in §7. Our method, which we will now present, looks at the

partial conjunction of the tests in combination with a sensitivity analysis.

3. Matched pair design with multiple treatments across pairs

3.1. Notation: K treatments in I pairs

There are I pairs of units matched on their observed covariates. Let ij, for $j = 1, 2$, index the units in pair i, $i = 1, \ldots, I$. The observed covariates for unit ij are $\mathbf{x}_{ij}$; $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ in each pair. Let $Z_{ij}^{(1)}$ be the indicator of exposure to treatment 1 for unit ij. In each pair there is one unit with treatment 1 and the other unit is not exposed to that treatment; so $Z_{i1}^{(1)} + Z_{i2}^{(1)} = 1$. Each unit is further exposed to treatments $2, \ldots, K$. We denote by $Z_{ij}^{(k)}$ the exposure status to treatment k for ij.

In the lead absorption study of §2, the first treatment, treatment 1, was employment of a parent in a battery manufacturing plant in Oklahoma. For an exposed child the subsequent treatments were based on parent's potential occupational exposure to lead — high or medium vs. low, treatment 2 and high vs. medium or low, treatment 3 — and further based on hygiene level of the parent — good or moderately good vs. poor, and good vs. moderately good or poor, treatment 4 and treatment 5 respectively. So, $I = 33$ and $K = 5$. Morton et al. collected data on occupation level of lead exposure and hygiene practice, only for the individuals exposed to treatment 1. Thus, the data for $Z_{ij}^{(2)}, \ldots, Z_{ij}^{(5)}$ were not available when $Z_{ij}^{(1)} = 0$. This does not hinder our analysis. As will become clear in our methodological development, the effect of treatment 2 will be analyzed only after conditioning on $Z_{ij} = 1$. Similarly, the effect of treatment 3 will be assessed only for exposed child with father exposed to high or medium level of occupational lead exposure. In practice, to create a matched design from two groups with $Z_i^{(1)} = 1$ and $Z^{(0)} = 0$, on a set of covariates $\mathbf{x}$, one can use algorithms available in the literature, see Hansen (2004); Pimentel et al. (2015) and Zubizarreta et al. (2014).

Let $Z_{ijk} = (Z_{ij}^{(1)}, \ldots, Z_{ij}^{(k)})$ be the k dimensional partial assignment vector of the first k treatments to unit ij, $1 \leq k \leq K$. The units are assumed to be assigned treatments independently — Z_{ijK} is independent of $Z_{i'j'K}$ for two different units ij and $i'j'$ across pairs, but the different treatments to a unit need not be assigned independently — $Z_{ij}^{(k)}$ need not be independent of $Z_{ij}^{(k')}$ for any k'. A father of an exposed child may have poor hygiene because he is accustomed to work in an environment where exposure to lead

is high, or he may have good hygiene. Since we make no assumption about the dependence structure of Z_{ijK}, either of the above associations is allowed in this model. Let $\mathbf{Z}_k = (Z_{11k}, Z_{12k}, \ldots, Z_{I2k})$ be the $2kI$ vector of first k treatment assignments on $2I$ units.

The outcome for unit ij is $R_{ij} = r_{ij}(Z_{ijK})$, determined from a set of 2^K potential outcomes, $r_{ij}(z_K)$ where $z_K \in \{0,1\}^K$ (Neyman, 1923; Rubin, 1974). Only a single element of this set is observed. If there is a causal effect, e.g. in §2, an effect of occupational exposure to lead, then the elaborate theory states that $r_{ij}(z_K) > r_{ij}(z'_K)$ for $z_K, z'_K \in \{0,1\}^K$ whenever $z_K \succ z'_K$ ($\succ$ denotes the partial ordering induced by coordinatewise ordering).

3.2. Assignment of treatment $\mathbf{Z}_K$

As mentioned above, it is assumed that Z_{ijK} is independent of $Z_{i'j'K}$ and that there is no interference between the units. This section defines the distribution of treatment exposure Z_{ijK}. The treatment assignment model is determined by the observed pre-treatment variables and the unmeasured confounders. This section also introduces the sensitivity parameters of our analysis.

Let $u_{ij1}, \ldots, u_{ijK}$ be K unmeasured variables, $0 \le u_{ijk} \le 1$, $1 \le k \le K$ (Rosenbaum, 2002). Set $\mathcal{F} = \{(\{r_{ij}(z_K), z_K \in \{0,1\}^K\}, \mathbf{x}_{ij}, u_{ij1}, \ldots, u_{ijK}); i = 1, \ldots, K, j = 1, 2\}$. We specify the distribution of Z_{ijK} as the product of conditional distributions, i.e. $\Pr(Z_{ijK} = z_{ijK} \mid \mathcal{F}) = \Pr(Z_{ij}^{(1)} = z_{ij}^{(1)} \mid \mathcal{F}) \prod_{k \ge 2} \Pr(Z_{ij}^{(k)} = z_{ij}^{(k)} \mid \mathcal{F}, Z_{ij(k-1)} = z_{ij(k-1)})$.

For the first treatment, treatment 1, we consider the model

$$\Pr(Z_{ij}^{(1)} = 1 \mid \mathcal{F}) = \frac{\exp(\theta_1(\mathbf{x}_{ij}) + \gamma_1 u_{ij1})}{1 + \exp(\theta_1(x_{ij}) + \gamma_1 u_{ij1})}. \tag{1}$$

Here, $\theta_1()$ is an arbitrary unknown function and $\gamma_1 \ge 0$ is a sensitivity parameter, also unknown. Under this model, as units are matched so that $Z_{i1}^{(1)} + Z_{i2}^{(1)} = 1$, we have

$$\Pr(Z_{i1}^{(1)} = 1 \mid \mathcal{F}, Z_{i1}^{(1)} + Z_{i2}^{(1)} = 1) = \frac{\exp(\gamma_1 u_{ij1})}{\exp(\gamma_1 u_{ij2}) + \exp(\gamma_1 u_{ij1})}. \tag{2}$$

With $\Gamma_1 = \exp(\gamma_1)$, the odds ratio of treatment 1 satisfies $\Gamma_1^{-1} \le \Pr(Z_{i1}^{(1)} = 1 \mid \mathcal{F}, Z_{i1}^{(1)} + Z_{i2}^{(1)} = 1) \Pr(Z_{i2}^{(1)} = 0 \mid \mathcal{F}, Z_{i1}^{(1)} + Z_{i2}^{(1)} = 1) \{\Pr(Z_{i1}^{(1)} = 0 \mid \mathcal{F}, Z_{i1}^{(1)} + Z_{i2}^{(1)} = 1) \Pr(Z_{i2}^{(1)} = 1 \mid \mathcal{F}, Z_{i1}^{(1)} + Z_{i2}^{(1)} = 1)\}^{-1} \le \Gamma_1$. When $\Gamma_1 = 1$ ($\gamma_1 = 0$) the odds ratio is 1 and the probability of unit ij getting treatment

1 in pair i is a coin flip. Thus, Γ_1 is a parameter that measures the deviation from the random assignment of treatment 1 in the pairs.

Consider the model for $Z_{ij}^{(k)}$ as

$$\Pr(Z_{ij}^{(k)} = 1 \mid \mathcal{F}, Z_{ij(k-1)} = z_{ij(k-1)}) = \frac{\exp(\theta_k(z_{ij(k-1)}) + \gamma_k u_{ijk})}{1 + \exp(\theta_k(z_{ij(k-1)}) + \gamma_k u_{ijk})}, \tag{3}$$

for $k \geq 2$. As before, $\theta_k()$ is an unknown function and $\gamma_k \geq 0$ is a sensitivity parameter.

Upon conditioning on $\mathbf{Z}_{k-1}$ the interpretation of γ_k becomes clearer when we consider the distribution of $(Z_{11}^{(k)}, Z_{12}^{(k)}, \ldots, Z_{I2}^{(k)})$. Let $\mathbf{a}_{k-1} \in \{0,1\}^{k-1}$, consider the set of all units with $Z_{ij(k-1)} = \mathbf{a}_{k-1}$; write it as $\mathcal{I}_{k-1}(\mathbf{a}_{k-1})$. Further write $|\mathcal{I}_{k-1}(\mathbf{a}_{k-1})| = n_{\mathbf{a}_{k-1}}$ for the number of these units. Denote by $Z^{(k)}(\mathcal{I}_{k-1}(\mathbf{a}_{k-1}))$ the vector of length $n_{\mathbf{a}_{k-1}}$ of kth treatment of the units in $\mathcal{I}_{k-1}(\mathbf{a}_{k-1})$ and by $u_k(\mathcal{I}_{k-1}(\mathbf{a}_{k-1}))$ the corresponding vector of kth unmeasured confounders, u_{ijk}'s. For $1 \leq m \leq n_{\mathbf{a}_{k-1}}$, let $\mathcal{Z}_{n_{\mathbf{a}_{k-1}},m}$ be the binary vectors of length $n_{\mathbf{a}_{k-1}}$ with m ones and $n_{\mathbf{a}_{k-1}} - m$ zeros. Then (3) implies

$$\begin{aligned} &\Pr(Z^{(k)}(\mathcal{I}_{k-1}(\mathbf{a}_{k-1})) = \mathbf{z} \mid \mathcal{F}, \mathbf{Z}_{k-1}, \sum_{ij \in \mathcal{I}_{k-1}(\mathbf{a}_{k-1})} Z_{ij}^{(k)} = m) \\ &= \frac{\exp(\gamma_k\, \mathbf{z}^\top u_k(\mathcal{I}_{k-1}(\mathbf{a}_{k-1})))}{\sum_{\zeta \in \mathcal{Z}_{n_{\mathbf{a}_{k-1}},m}} \exp(\gamma_k\, \zeta^\top u_k(\mathcal{I}_{k-1}(\mathbf{a}_{k-1})))}, \quad \text{for } \mathbf{z} \in \mathcal{Z}_{n_{\mathbf{a}_{k-1}},m}. \end{aligned} \tag{4}$$

Irrespective of the value of u_{ijk}'s, if $\gamma_k = 0$ ($\Gamma_k := \exp(\gamma_k) = 1$), this probability is $\binom{n_{\mathbf{a}_{k-1}}}{m}^{-1}$, which indicates a randomized assignment of m units to be treated with treatment k among the units in $\mathcal{I}(\mathbf{a}_{k-1})$. The larger the value of Γ_k is the bias in treatment k is further from this random assignment.

Remark. Models (1) and (3) are our sensitivity analysis models. The parameters Γ_1 and Γ_k's are the sensitivity parameters whose values we choose to constrain the amount of bias in the treatment assignment due to unmeasured confounding. Further, these models are also fully nonparametric, in the following sense. If Γ_1 is the bias in treatment 1, so that for any two units which are similar in their observed covariates, the odds ratio of being exposed to treatment 1 is at most Γ_1, then, there exists θ_1 and u_{ij1}'s so that (1) holds. For a proof of this statement see Rosenbaum (2002), §4.2. Similarly, a specification of Γ_k is equivalent to model (3).

3.3. K tests for the causal hypothesis and their sensitivity to unmeasured confounding

The causal hypothesis has broad implications. When it is true, an exposure to the treatment, at any level, increases the outcome. This section reviews various nonparametric test statistics for the implications of the causal hypothesis and, using the treatment assignment model discussed in §3.2, also reviews the methods to assess the sensitivity of these tests to unmeasured confounders. Consider ranking of the responses by a preferred choice of ranking/scoring method for the K tests. Let q_{ijk} be the nonnegative score of unit ij for test k, $k = 1, \ldots, K$. The scores are determined from the observed outcomes $(R_{11}, R_{12}, \ldots, R_{I2})$.

Fix $\mathbf{a} = (a_1, \ldots, a_K) \in \{0,1\}^K$ and let $\mathbf{a}_{k-1} = (a_1, \ldots, a_{k-1})$, $2 \le k \le K$. For convenience we further write for $k = 1$, $k-1 = 0$, $\mathbf{a}_{k-1} = \mathbf{a}_0 = \emptyset$. As in our discussion of §3.2, let $\mathcal{I}_{k-1}(\mathbf{a}_{k-1})$ be the set of units with $Z_{ij(k-1)} = \mathbf{a}_{k-1}$. Set $\mathcal{I}_0(\mathbf{a}_0) = \mathcal{I}_0(\emptyset)$ to be the set of all $2I$ study units. Then we consider the following form the test statistics for the paired comparison on treatment 1

$$T_{1,\mathbf{a}_0} = \sum_{i=1}^{I} \mathrm{sgn}\{(Z_{i1}^{(1)} - Z_{i2}^{(1)})(R_{i1} - R_{i2})\}(q_{i11} + q_{i21}).$$

The function $\mathtt{sgn}(x)$ is -1, 0 or 1 depending on $x < 0$, $x = 0$ or $x > 0$. Our test statistics for the effect of treatment $k \ge 2$ is

$$T_{k,\mathbf{a}_{k-1}} = \sum_{ij \in \mathcal{I}_{k-1}(\mathbf{a}_{k-1})} Z_{ij}^{(k)} q_{ijk}.$$

When $k = 1$ the test statistics is a pairwise comparison. In particular, if $q_{i1k} = q_{i2k}$ is the rank of absolute difference $|R_{i1} - R_{i2}|$ in the sorted list of the pairwise absolute differences, then $T_{1,\mathbf{a}_0}$ is twice the Wilcoxon signed rank test statistics. When $k \ge 2$ the test is across pairs. But since it conditions on $\mathbf{a}_{k-1}$, thus in particular fixes $Z_{ij}^{(1)}$ of all the units in $\mathcal{I}_{k-1}(\mathbf{a}_{k-1})$, at most one unit from each pair is considered. Technically though there is no harm in scoring $ij \in \mathcal{I}_{k-1}(\mathbf{a}_{k-1})$ as q_{ijk} by also using outcomes of units $i'j' \notin \mathcal{I}_{k-1}(\mathbf{a}_{k-1})$.

Let $P_{k,\mathbf{a}_{k-1}}$ be the p-value assessing the extent to which the test statistics $T_{k,\mathbf{a}_{k-1}}$ provides evidence for an effect of treatment k. The null hypothesis, H_0, is Fisher's sharp null so that $r_{ij}(z_{ijK}) = r_{ij}(z'_{ijK})$ for all ij and $z_{ijK}, z'_{ijK} \in \{0,1\}^K$. If $T^{obs}_{k,\mathbf{a}_{k-1}}$ is the observed value of the test statistics in the data then

$$P_{k,\mathbf{a}_{k-1}} = \Pr(T_{k,\mathbf{a}_{k-1}} \ge T^{obs}_{k,\mathbf{a}_{k-1}} \mid \mathcal{F}, \mathbf{Z}_{k-1}, \sum_{ij \in \mathcal{I}_{k-1}(\mathbf{a}_{k-1})} Z_{ij}^{(k)}, H_0). \tag{5}$$

The test for the effect of exposure to kth treatment conditions on $\mathbf{Z}_{k-1}$ and $\sum_{ij\in\mathcal{I}_{k-1}(\mathbf{a}_{k-1})} Z_{ij}^{(k)}$ as they are irrelevant for the effect (Kalbfleish, 1975; Heland, 1995). Conditioning on H_0 does not affect the treatment assignment distributions (1)–(4). If we could know u_{ijk}, we would calculate these p-values from the first principle using the probability distribution (2) if $k = 1$ and (4) if $k \geq 2$. The same is true if $\gamma_k = 0$. In the former of these two cases there is potentially bias from confounding variable, but these variables are known. In the second scenario there is no bias from unmeasured confounding and we use the conditional randomization distribution of the treatment k for calculating the p-values.

However, the unmeasured confounders, u_{ijk}'s are just that — unmeasured. Thus, $P_{k,\mathbf{a}_{k-1}}$ cannot be calculated if $\gamma_k > 0$. We calculate the maximum value of the p-value $P_{k,\mathbf{a}_{k-1}}$, after fixing $\Gamma_k = \exp(\gamma_k)$, over the range of u_{ijk}; call this maximum $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k}$. The calculation is different between $\overline{P}_{1,\mathbf{a}_0,\Gamma_1}$, the paired comparison for treatment 1, and $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k}$ for $k \geq 2$, between pair comparisons. Consider the paired comparison. Then

$$\overline{P}_{1,\mathbf{a}_0,\Gamma_1} = \Pr(\sum_{i=1}^{I} s_i(q_{i11} + q_{i21}) \geq T_{1,\mathbf{a}_0}^{obs} \mid \mathcal{F}),$$

where s_i's are independently distributed taking values 1 with probability $\Gamma_1/(1+\Gamma_1)$ and -1 with probability $(1+\Gamma_1)^{-1}$ if $R_{i1} \neq R_{i2}$ and $s_i \equiv 0$ if $R_{i1} = R_{i2}$ (Rosenbaum, 1987; 2002, §4.3).

The finite sample calculation of $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k}$, $k \geq 2$, is cumbersome. Recall $\mathcal{I}_{k-1}(\mathbf{a}_{k-1})$ is the set of units with $Z_{ij(k-1)} = \mathbf{a}_{k-1}$. Let $n_{\mathbf{a}_{k-1}} = |\mathcal{I}_{k-1}(\mathbf{a}_{k-1})|$ and $m = \sum_{ij\in\mathcal{I}_{k-1}(\mathbf{a}_{k-1})} Z_{ij}^{(k)}$. Temporarily denote the units in $\mathcal{I}_{k-1}(\mathbf{a}_{k-1})$ by $\tilde{i}1, \ldots, \tilde{i}n_{\mathbf{a}_{k-1}}$ so that the corresponding k scores are sorted in increasing order, $q_{\tilde{i}1k} \leq \cdots \leq q_{\tilde{i}n_{\mathbf{a}_{k-1}}k}$. There are $2^{n_{\mathbf{a}_{k-1}}}$ values of $u(\mathcal{I}_{k-1}(\mathbf{a}_{k-1}))$ to maximize over. This can immediately be reduced to maximizing over only $n_{i,\mathbf{a}_{k-1}} - 1$ of them. These $u(\mathcal{I}_{k-1}(\mathbf{a}_{k-1}))$'s correspond to an $l = 1, \ldots, n_{i,\mathbf{a}_{k-1}} - 1$, so that $u_{\tilde{i}1k} = \cdots = u_{\tilde{i}lk} = 0$ and $u_{\tilde{i}(l+1)k} = \cdots = u_{\tilde{i}n_{\mathbf{a}_{k-1}}k} = 1$. Still, the exact evaluation of the probabilities for these l instances is less than efficient. We consider the large sample approximation bound. It requires the following function

$$\mathcal{C}_k(a,b,c) = \sum_{l=\max(0,b+c-a)}^{\min(b,c)} \binom{c}{l}\binom{a-c}{b-l} e^{\gamma_k l} \cdot \mathbf{1}(a \geq b, b > 0, c > 0). \tag{6}$$

This function was discussed in Rosenbaum and Krieger (1990), equation (8). Let $\Sigma_{l,\mathbf{a}_{k-1}}$ be a symmetric matrix of size $n_{\mathbf{a}_{k-1}}$ defined as follows. The diago-

nal element of this matrix is $\Sigma_{l,\mathbf{a}_{k-1}}(\tilde{j},\tilde{j}) = \mathcal{C}_k(n_{\mathbf{a}_{k-1}}-1, m-1, l)\{\mathcal{C}_k(n_{\mathbf{a}_{k-1}}, m, l)\}^{-1}$ if $\tilde{j} \leq l$ and $\Sigma_{l,\mathbf{a}_{k-1}}(\tilde{j},\tilde{j}) = \Gamma_k\mathcal{C}_k(n_{\mathbf{a}_{k-1}}-1, m-1, l-1)\{\mathcal{C}_k(n_{\mathbf{a}_{k-1}}, m, l)\}^{-1}$ if $\tilde{j} \geq l+1$. The $(\tilde{j},\tilde{j}')$th off-diagonal element of this symmetric matrix is $\mathcal{C}_k(n_{\mathbf{a}_{k-1}}-2, m-2, l)\{\mathcal{C}_k(n_{\mathbf{a}_{k-1}}, m, l)\}^{-1}$ if $\tilde{j} \leq l$ and $\tilde{j}' \leq l$; it is $\Gamma_k\mathcal{C}_k(n_{\mathbf{a}_{k-1}}-2, m-2, l-1)\{\mathcal{C}_k(n_{\mathbf{a}_{k-1}}, m, l)\}^{-1}$ if $\tilde{j} \leq l$ and $\tilde{j}' \geq l+1$; and it is $\Gamma_k^2\mathcal{C}_k(n_{\mathbf{a}_{k-1}}-2, m-2, l-2)\{\mathcal{C}_k(n_{\mathbf{a}_{k-1}}, m, l)\}^{-1}$ if $\tilde{j} \geq l+1$ and $\tilde{j}' \geq l+1$. Then the mean of the test statistics for the unmeasured confounder l is

$$\begin{aligned}\mu_{l,\mathbf{a}_{k-1}} &= \sum_{\tilde{j}=1}^{l} \frac{\mathcal{C}_k(n_{\mathbf{a}_{k-1}}-1, m-1, l)}{\mathcal{C}_k(n_{\mathbf{a}_{k-1}}, m, l)} q_{\tilde{i}\tilde{j}k} + \sum_{\tilde{j}=l+1}^{n_{\mathbf{a}_{k-1}}} \Gamma_k \frac{\mathcal{C}_k(n_{\mathbf{a}_{k-1}}-1, m-1, l-1)}{\mathcal{C}_k(n_{\mathbf{a}_{k-1}}, m, l)} q_{\tilde{i}\tilde{j}k} \\ &= \sum_{\tilde{j}=1}^{n_{\mathbf{a}_{k-1}}} \Sigma_{l,\mathbf{a}_{k-1}}(\tilde{j},\tilde{j}) q_{\tilde{i}\tilde{j}k},\end{aligned}$$

and the variance is

$$\nu^2_{l,\mathbf{a}_{k-1}} = \sum_{\tilde{j},\tilde{j}'=1}^{n_{\mathbf{a}_{k-1}}} \Sigma_{l,\mathbf{a}_{k-1}}(\tilde{j},\tilde{j}') q_{\tilde{i}\tilde{j}k} q_{\tilde{i}\tilde{j}'k} - \left(\mu_{l,\mathbf{a}_{k-1}}\right)^2.$$

Then the asymptotically correct value, as $I \to \infty$, of the maximum p-value for the kth test statistics is (Rosenbaum, 2002, §4.6, §4.7)

$$\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} = 1 - \min_{l=1,\ldots,n_{\mathbf{a}_{k-1}}-1} \Phi^{-1}((T^{obs}_{k,\mathbf{a}_{k-1}} - \mu_{l,\mathbf{a}_{k-1}})/\nu_{l,\mathbf{a}_{k-1}}).$$

For each l, the computation of $\mu_{l,\mathbf{a}_{k-1}}$ is a multiplication of two vectors of size $n_{\mathbf{a}_{k-1}}$. The computation of $\nu_{l,\mathbf{a}_{k-1}}$ requires calculation of a quadratic form for a square matrix of size $n_{\mathbf{a}_{k-1}}$. Thus, when the values of the function $\mathcal{C}_k(a,b,c)$ can be queried in constant time, the calculation of the means and the variances together has a computational complexity of $O(n^2_{\mathbf{a}_{k-1}})$. To implement the proposed methods, it would make sense to have the values of the function computed beforehand and stored, since they do not require the data. Also, the computational cost of these functions is at most $O(m) = O(n_{\mathbf{a}_{k-1}})$ when the coefficients in the summands of (6), the products of the binomial coefficients, are pre-stored. Each l only requires 6 values of this function to define $\Sigma_{l,\mathbf{a}_{k-1}}$. The method for computing the bounds $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k}$ has been implemented in the `R` package `senstrat`.

For various methods of sensitivity analysis in observational studies, see Cornfield et al. (1959), Egleston et al. (2009), Fogarty ans Small (2016), Fogarty and Hasegawa (2018), Gilbert et al. (2010), Hosman et al. (2010), Liu et al. (2013), and Yu and Gastwirth (2005). In particular see Rosenbaum (2018) for a comprehensive discussion and faster computation of $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k}$.

4. Evidence factors and pooling evidence

In the lead absorption study of §2, there are $K = 5$ tests with $\mathbf{a} = (1, 1, 1, 1, 1)$ or $\mathbf{a} = (1, 1, 1, 1, 0)$; the last coordinate is irrelevant for the design of the tests. The K test statistics are $T_{k,\mathbf{a}_{k-1}}$, $1 \le k \le K$. The previous section showed the computation of the maximum p-values for the these test statistics when the bias from unmeasured confounders is at most Γ_k. These maximum p-values are denoted by $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k}$. Considered separately, for significance level α, the test using $T_{k,\mathbf{a}_{k-1}}$ is sensitive at level Γ_k if $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \ge \alpha$. This section establishes that these tests form evidence factors — they are biased by separate confoundings and they are nearly independent when the null is true.

Proposition 1 *Fix $\mathbf{a} \in \{0,1\}^K$. Under H_0, when the treatment assignment model is as (1) and (3), i.e. the bias in treatment k is at most $\Gamma_k = \exp(\gamma_k)$*

$$\Pr(\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \le \alpha_k \, \forall k \ge 1 \mid \mathcal{F}) \le \prod_{k=1}^{K} \alpha_k.$$

Proof. We first note that $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k}$, which is the maximum value of $P_{k,\mathbf{a}_{k-1},\Gamma_k}$ in (5) under model, is a function of $\mathcal{F}$, $\mathbf{Z}_{k-1}$ and $\sum_{ij \in \mathcal{I}_{k-1}(\mathbf{a}_{k-1})} Z_{ij}^{(k)}$. We write

$$\begin{aligned}
&\Pr(\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \le \alpha_k \, \forall k \ge 1 \mid \mathcal{F}) \\
&= \Pr(\overline{P}_{1,\mathbf{a}_0,\Gamma_1} \le \alpha_1 \mid \mathcal{F}) \times \prod_{k=2}^{K} \Pr(\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \le \alpha_k \mid \overline{P}_{k',\mathbf{a}_{k'-1},\Gamma_{k'}} \le \alpha_{k'} \, \forall k' \le k-1, \mathcal{F}).
\end{aligned}$$

Under H_0 and (1), $\Pr(\overline{P}_{1,\mathbf{a}_0,\Gamma_1} \le \alpha_1 \mid \mathcal{F}) \le \alpha_1$. Further for any $k \ge 2$, $\Pr(\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \le \alpha_k \mid \overline{P}_{k',\mathbf{a}_{k'-1},\Gamma_{k'}} \le \alpha_{k'} \, \forall k' \le k-1, \mathcal{F}) = E[E\{\mathbf{1}(\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \le \alpha_k) \mid \overline{P}_{k',\mathbf{a}_{k'-1},\Gamma_{k'}} \le \alpha_{k'} \, \forall k' \le k-1, \mathcal{F}, \mathbf{Z}_{k-1}, \sum_{ij \in \mathcal{I}_{k-1}(\mathbf{a}_{k-1})} Z_{ij}^{(k)}\}]$. The outer expectation marginalizes over $\mathbf{Z}_{k-1}$ and $\sum_{ij \in \mathcal{I}_{k-1}(\mathbf{a}_{k-1})} Z_{ij}^{(k)}$. Under H_0 and (3), by (5), the inner expectation is at most α_k. Combining these facts gives the required result. ■

Theorem 1 *Fix $\mathbf{a} \in \{0,1\}^K$. Let $f : [0,1]^K \to (-\infty, \infty)$ be a function which is nondecreasing in its coordinates, i.e. $f(x_1, \ldots, x_k, \ldots, x_K) \ge f(x_1, \ldots, x_k', \ldots, x_K)$ for any $x_k' \ge x_k$. Suppose $U_1, \ldots, U_K$ are K i.i.d. random variables uniformly distributed on $[0,1]$. Under H_0, when the treatment assignment model is as in (1) and (3), for $-\infty \le x \le \infty$,*

$$\Pr(f(\overline{P}_{1,\mathbf{a}_0,\Gamma_1}, \ldots, \overline{P}_{K,\mathbf{a}_{K-1},\Gamma_K}) \le x \mid \mathcal{F}) \le \Pr(f(U_1, \ldots, U_K) \le x).$$

Proof. The proof of the theorem follows from Proposition 1, along with Theorem 6.B.4 and Theorem 6.B.16 of Shaked and Shanthikumar (2007). A more general statement, Theorem 2, is proved in the appendix. ■

Theorem 1 shows that the joint distribution of the K p-values is stochastically larger than the uniform distribution on K dimensional hyper-cube. Thus, the tests are nearly independent in the sense of Theorem 1. Thus, in the lead study, the maximum p-values corresponding to testing the $K = 5$ pieces of the elaborate theory are nearly independent. The consequence of Theorem 1 is that usual methods of combining independent p-values can be used to pool evidence and report a single number for the evidence against the null that there is no causal effect. In particular, one can use Fisher's method (Fisher, 1932) of combining p-values to calculate $P_K^{Fisher} = \Pr(\chi^2_{2K} > -2\sum_{k=1}^{K} \log \overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k})$. The dependence of P_K^{Fisher} on Γ_k's is suppressed here for convenience of notation. Theorem 1 implies that for any $\alpha \in [0,1]$, when the biases are at most Γ_k, under H_0, $\Pr(P_K^{Fisher} \leq \alpha) \leq \alpha$. There are many such methods. Becker (1994) is a convenient reference for such methods. Zaykin et al. (2002)s' method deserves special mention. Zaykin et al. proposed a variant of the Fisher's method by combining independent p-values using a truncated product. The test statistics is a product of those p-values that are smaller than some truncation point, $\varkappa$. Hsu et al. (2013) show that the truncated product with $\varkappa = 0.20$ or $\varkappa = 0.10$ often has higher power than Fisher's method when applied to p-value bounds from a sensitivity analysis. The intuition is: the individual maximum p-values are not uniform but rather stochastically larger than a uniform distribution on $[0,1]$, thus conservative.

5. Evidence from a partial conjunction of the tests: A quantification of the extent of corroboration

The pooled evidence from all the K tests has the benefit of ease of interpretation, yet it only provide information on whether at least one of the K tests support the alternative hypothesis, not whether a larger fraction support the alternative hypothesis. This section considers evidence from partial conjunctions of the tests. Throughout this section we fix $\mathbf{a} \in \{0,1\}^K$.

Fix k, $1 \leq k \leq K$. The null hypothesis for the effect of treatment k is the hypothesis that treatment k does not change the potential outcome of the units. Written formally $H_{0,k} : R_{ij}(z_{ijK}) = R_{ij}(z'_{ijK})$ for $z_{ijK}, z'_{ijK} \in \{0,1\}^K$ if ${z_{ijK}}^{(k')} = {z'_{ijK}}^{(k')}$ for all $k' \leq k-1$; the alternative, $H_{1,k}$, states that treatment k increases the response. The test statistics $T_{k,\mathbf{a}_{k-1}}$ tests for this null

hypothesis. The global null H_0 is equivalent to $\cap_{k=1}^{K} H_{0,k}$. Indeed, in (5) we can replace H_0 by $H_{0,k}$, all arguments of §3.3 and §4 remain unchanged. The pooled evidence as in §4 is evidence against intersection of K nulls $H_{0,k}$s. A small value of the pooled evidence tells us that we have evidence for at least one of these (one sided) alternatives. Consequently, it preserves the familywise error rate: "Pr(Reject at least one $H_{0,k}; k = 1, \ldots, K) \leq$ Pr(Reject $\cap_{k=1}^{K} H_{0,k}$) = Pr(Reject H_0)".

The global null H_0 is still false if at least one of the hypotheses is false, or at least k of them are false. Is there evidence that at least k of the K hypotheses are false? Is there evidence for the causal hypothesis that occupational exposure to lead among parents causes childrens' lead level to increase based on the k of the $K = 5$ pieces of the elaborate theory? Write, for $1 \leq k \leq K$

$$H_0^{k|K} : \cup_{l=K-k+1}^{K} \cap_{t \in \{t_1,\ldots,t_l\}, 1 \leq t_1 < \cdots < t_l \leq K} H_{0,t},$$

for the hypothesis that at most $k-1$ of the K nulls are false. If $H_0^{k|K}$ is false then at least k hypotheses are false. Specifically, $H_0^{1|K} \equiv H_0$. The evidence against $H_0^{k|K}$, i.e. evidence that at least k of the null hypotheses are false, is found by looking at the largest $K - k + 1$ p-values. Recall the p-values bounds were denoted by $(\overline{P}_{1,\mathbf{a}_0,\Gamma_1}, \ldots, \overline{P}_{K,\mathbf{a}_{K-1},\Gamma_K})$. Let $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_K)$. We denote by $\overline{P}_{(1)\mathbf{a},\mathbf{\Gamma}} \leq \cdots \leq \overline{P}_{(K)\mathbf{a},\mathbf{\Gamma}}$, those K values in increasing order. Consider a function $g_k : [0,1]^{K-k+1} \to [0,1]$. Then the evidence against $H_0^{k|K}$ has the form

$$P_{\mathbf{a},\mathbf{\Gamma}}^{k|K} = g_k(\overline{P}_{(k)\mathbf{a},\mathbf{\Gamma}}, \ldots, \overline{P}_{(K)\mathbf{a},\mathbf{\Gamma}}). \tag{7}$$

Theorem 2 is a general statement of Proposition 1 and Theorem 1 for any subset of the tests. The proof of Theorem 2 is given in the appendix. This theorem will be required to study the p-values $P_{\mathbf{a},\mathbf{\Gamma}}^{k|K}$s.

Theorem 2 *Fix* $\mathbf{a} \in \{0,1\}^K$. *Let* $\mathcal{K} = \{k_1, \ldots, k_{|\mathcal{K}|}\} \subseteq \{1, \ldots, K\}$. *Under* $\cap_{t \in \mathcal{K}} H_{0,t}$, *when treatment assignment model is* (1) *and* (3), *but only for* $k \in \mathcal{K}$, *then for any nondecreasing function* $f_{\mathcal{K}} : [0,1]^{|\mathcal{K}|} \to (-\infty, \infty)$, *for* $|\mathcal{K}|$ *i.i.d. uniform [0,1] random variables* $U_1, \ldots, U_{|\mathcal{K}|}$, *and* $-\infty < x < \infty$,

$$\Pr(f_{\mathcal{K}}(\overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}}, \ldots, \overline{P}_{k_{|\mathcal{K}|},\mathbf{a}_{k_{|\mathcal{K}|}-1},\Gamma_{k_{|\mathcal{K}|}}}) \leq x \mid \mathcal{F}) \leq \Pr(f_{\mathcal{K}}(U_1, \ldots, U_{|\mathcal{K}|}) \leq x).$$

The kth test become sensitive for bias level Γ_k when $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \geq \alpha$. The test for the partial conjunction hypothesis, $H_0^{k|K}$, is sensitive at bias level $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_k)$ if the pooled p-value is more than α, $P_{\mathbf{a},\mathbf{\Gamma}}^{k|K} \geq \alpha$. Using

Theorem 2 the following proposition establishes that $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}$ in (7) is a p-value for testing $H^{k|K}_0$. Proposition 2 is equivalent to Theorem 1 of Benjamini and Heller (2008). See also Wang and Owen (2017) for related results.

Proposition 2 *Fix* $\mathbf{a} \in \{0,1\}^K$. *Consider model* (1) *and* (3). *Let* $g_k : [0,1]^{K-k+1} \to [0,1]$ *be a coordinatewise nondecreasing function in* (7). *Suppose,* $\Pr(g_k(U_k,\ldots,U_K) \le \alpha) \le \alpha$ *for some* $\alpha \in [0,1]$, *where* $U_1,\ldots,U_K$ *are i.i.d uniform random variables on* $[0,1]$. *Then, under* $H^{k|K}_0$

$$\Pr(P^{k|K}_{\mathbf{a},\mathbf{\Gamma}} \le \alpha \mid \mathcal{F}) \le \alpha. \tag{8}$$

Proof. Recall, $H^{k|K}_0 : \cup^K_{l=K-k+1} \cap_{t\in\{t_1,\ldots,t_l\},1\le t_1<\cdots<t_l\le K} H_{0,t}$. Fix, $1 \le t_1 < \cdots < t_l \le K$ for some $l \ge K-k+1$ and set $\mathcal{K} = \{t_1,\ldots,t_{K-k+1}\}$. Then, $\cap_{t\in\{t_1,\ldots,t_l\}} H_{0,t}$ implies $\cap_{t\in\mathcal{K}} H_{0,k}$. By (7), with the fact that g_k is coordinatewise nondecreasing and Theorem 2, respectively, we bound the probability in (8) by

$$\Pr(g_k(\overline{P}_{t_1,\mathbf{a}_{t_1-1},\Gamma_{t_1}},\ldots,\overline{P}_{t_{K-k+1},\mathbf{a}_{t_{K-k+1}-1},\Gamma_{t_{K-k+1}}}) \le \alpha \mid \mathcal{F}) \le \Pr(g_k(U_1,\ldots,U_{K-k+1}) \le \alpha) \le \alpha.$$

■

Proposition 3 *Consider K functions,* $g_k : [0,1]^{K-k+1} \to [0,1]$, $1 \le k \le K$. *Assume the following, for i.i.d. uniform [0,1] random variables* $U_1,\ldots,U_K$, *for all* $k = 1,\ldots,K$

(a) g_k *is nondecreasing in its coordinates.*
(b) $\Pr(g_k(U_k,\ldots,U_K) \le \alpha) \le \alpha$, *for some* $\alpha \in [0,1]$.
(c) $g_k(x_k,x_{k+1},\ldots,x_K) \le g_{k+1}(x_{k+1},\ldots,x_K)$ *for all* $x_{k+1},\ldots,x_K \in [0,1]$ *and* $x_k \le \min\{x_{k+1},\ldots,x_K\}$.

Condition (c) is void if $k = K$. *Fix* $\mathbf{a} \in \{0,1\}^K$. *Suppose we reject* $H^{k|K}_0$ *if* $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}} = g_k(\overline{P}_{(k)\mathbf{a},\mathbf{\Gamma}},\ldots,\overline{P}_{(K)\mathbf{a},\mathbf{\Gamma}})$ *is less than* α. *Under model* (1) *and* (3), *the probability of rejecting any true null hypothesis among* $\{H^{k|K}_0; k = 1,\ldots,K\}$ *is at most* α.

Proof. Since, $H^{k|K}_0$ is the hypothesis that at most $k-1$ nulls are false, they satisfy $H^{1|K}_0 \subseteq \cdots \subseteq H^{K|K}_0$. Further, condition (c) implies $P^{1|K}_{\mathbf{a},\mathbf{\Gamma}} \le \cdots \le P^{K|K}_{\mathbf{a},\mathbf{\Gamma}}$. This is because, by (c), for $k = 1,\ldots,K-1$, $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}} = g_k(\overline{P}_{(k)\mathbf{a},\mathbf{\Gamma}},\ldots,\overline{P}_{(K)\mathbf{a},\mathbf{\Gamma}}) \le g_{k+1}(\overline{P}_{(k+1)\mathbf{a},\mathbf{\Gamma}},\ldots,\overline{P}_{(K)\mathbf{a},\mathbf{\Gamma}}) = P^{k+1|K}_{\mathbf{a},\mathbf{\Gamma}}$.

If there is no true null among $\{H^{k|K}_0; k = 1,\ldots,K\}$ there is nothing to prove. Otherwise, let k be the smallest number such that $H^{k|K}_0$ is

true. Consequently, $H_0^{1|K}, \ldots, H_0^{k-1|K}$ are false. Then a false rejection implies rejection of a null hypothesis $H_0^{k'|K}$ which is true and $k' \geq k$ with $P_{\mathbf{a},\mathbf{\Gamma}}^{k'|K} < \alpha$. From the ordering of the p-values noted above, it implies $P_{\mathbf{a},\mathbf{\Gamma}}^{k|K} < \alpha$. Hence the probability of rejecting any true null hypothesis among $\{H_0^{k|K}; k = 1, \ldots, K\}$ is bounded by $\Pr(P_{\mathbf{a},\mathbf{\Gamma}}^{k|K} < \alpha \mid \mathcal{F}, H_0^{k|K})$. This is at most α by condition (a) and (b) using Proposition 2. ■

By the above proposition, for the proposed method, for testing the set of K hypotheses for the partial conjunctions of the different pieces of the elaborate theory, the type-I error rate is at most the nominal level α.

Condition (c) of Proposition 3 is satisfied by Simes' method of combining p-values (Simes, 1986). To see this, consider $0 \leq x_k \leq x_{k+1} \leq \cdots \leq x_K \leq 1$. Simes' method uses the function $g_k(x_k, x_{k+1}, \ldots, x_K) = \min_{l=1,\ldots,K-k+1} l^{-1}(K-k+1)x_{k+l-1}$ in calculating $P_{\mathbf{a},\mathbf{\Gamma}}^{k|K}$ using (7). Accordingly, $g_{k+1}(x_{k+1}, \ldots, x_K) = \min_{l=1,\ldots,K-k} l^{-1}(K-k)x_{k+l} = \min_{l=2,\ldots,K-k+1}(l-1)^{-1}(K-k)x_{k+l-1}$. It follows that

$$\begin{aligned}
g_k(x_k, x_{k+1}, \ldots, x_K) &= \min_{l=1,\ldots,K-k+1} l^{-1}(K-k+1)x_{k+l-1} \\
&\leq \min_{l=2,\ldots,K-k+1} l^{-1}(K-k+1)x_{k+l-1} \\
&= \min_{l=2,\ldots,K-k+1} \{(l-1)l^{-1}(K-k+1)(K-k)^{-1}\}(l-1)^{-1}(K-k)x_{k+l-1} \\
&\leq \min_{l=2,\ldots,K-k+1} (l-1)^{-1}(K-k)x_{k+l-1} \\
&= g_{k+1}(x_{k+1}, \ldots, x_K).
\end{aligned}$$

Although, this condition may not be satisfied generally by any method of combining p-values. For example, it is not satisfied by Fisher's method. To see this let $K = 2$, $x_1 = x_2 = 0.5$. Then $g_1(x_1, x_2) = \Pr(\chi_4^2 > -2\log x_1 \cdot x_2) \approx 0.596 > 0.5 = \Pr(\chi_2^2 > -2\log x_2) = g_2(x_2)$. The following proposition lists other methods that satisfies the conditions (a)–(c) of Proposition 3. The first one in this list looks only at the minimum p-value $\overline{P}_{(k)\mathbf{a},\mathbf{\Gamma}}$ for testing $H_0^{k|K}$. This 'minimum p-value' method is fairly well known in the statistics literature. The following method is Stouffer's method which is popular in the meta-analysis literature (Stouffer et al., 1949). The last method in this list is a modification of 'additive p-value method' of Edgington (1972).

Proposition 4 *Conditions (a)–(c) of Proposition 3 are satisfied by each of the following specifications of g_ks.*

1. *(minimum p-value method)* $g_k(x_k, \ldots, x_K) = 1-(1-\min\{x_k, \ldots, x_K\})^{K-k+1}$.

2. *(sum of z's)* $g_k(x_k, \ldots, x_K) = 1 - \Phi(\,(\Phi^{-1}(1-x_k) + \cdots + \Phi^{-1}(1-x_K))/\sqrt{(K-k+1)}\,)$.
3. *(modified additive p-value method) With* $A_k = x_k + \cdots + x_K$, $g_k(x_k, \ldots, x_K) = (\min\{\frac{A_k^{K-k+1}}{(K-k+1)!}, 1\})^{\mathbf{1}(A_k \le c_k)}$ *where* $c_k = (K-k+1)(1-(K-k+2)^{-1})^{K-k+1}$.

The proof of this proposition is given in the appendix. It might often be useful to weight the p-values when combining them. However, the validity of the combined p-value for the partial conjunction hypothesis would usually require the weights to be predetermined. Also, the optimal choice of the weights could depend on the specific problem (Chen, 2011; Lancaster, 1961; Lipák, 1958; Whitlock, 2005; Zaykin, 2011). We do not discuss the various methods of weighted combinations in this paper.

The rest of this section considers the sensitivity analysis to unmeasured confounding over the multiple sensitivity parameters. There are K sensitivity parameters, $\Gamma_1, \ldots, \Gamma_K$. We gradually establish that the proposed sensitivity analysis for testing of partial conjunction of the hypotheses will control for the familywise error rate. These results ensure the validity of our analysis, which is presented in §7, of the elaborate theory of the causal hypothesis for the effect of occupational lead exposure among parents on children.

In the sensitivity analysis one first fixes a range of values of the bias parameters. Let $1 = \Gamma_{11} < \cdots < \Gamma_{1S_1}$ be the range of values for the bias parameter Γ_1 for bias in treatment 1; $1 = \Gamma_{k1} < \cdots < \Gamma_{kS_k}$ is the range of values for the bias parameter Γ_k for treatment k. Let $\mathfrak{J} = \{\mathbf{\Gamma} = (\Gamma_{1s_1}, \ldots, \Gamma_{Ks_K}) : 1 \le s_1 \le S_1; \cdots; 1 \le s_K \le S_K\}$. The goal is to find the least amount of bias that could explain an observed association. We denote by $H_{0,\mathbf{\Gamma}}^{k|K}$ the conjunction of the hypothesis $H_0^{k|K}$ and that the bias is at most $\mathbf{\Gamma}$. The statement that — the bias is at most $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_K)$ — means the treatment assignment satisfies (1) and (3) with $\gamma_k = \log \Gamma_k$ for some set of unmeasured confounders u_{ijk}'s. The following theorem says that the maximum error of the multi-parameter sensitivity analysis using $P_{\mathbf{a},\mathbf{\Gamma}}^{k|K}$s is bounded by α.

Theorem 3 *Fix* k, $1 \le k \le K$. *Consider the set of sensitivity parameters* $\mathfrak{J} = \{\mathbf{\Gamma} = (\Gamma_{1s_1}, \ldots, \Gamma_{Ks_K}) : 1 \le s_1 \le S_1; \cdots; 1 \le s_K \le S_K\}$. *Assume the conditions of Proposition 2. Fix* $\mathbf{a} \in \{0,1\}^K$. *Consider the procedure that rejects* $H_{0,\mathbf{\Gamma}}^{k|K}$ *for* $\mathbf{\Gamma} \in \mathfrak{J}$ *if* $P_{\mathbf{a},\mathbf{\Gamma}}^{k|K} < \alpha$. *Then the probability of rejecting any true null hypothesis among the set of hypotheses* $\{H_{0,\mathbf{\Gamma}}^{k|K}; \mathbf{\Gamma} \in \mathfrak{J}\}$ *is at most* α.

Proof. Note first that $H^{k|K}_{0,\mathbf{\Gamma}} \subseteq H^{k|K}_{0,\mathbf{\Gamma}'}$ for $\mathbf{\Gamma}' \succ \mathbf{\Gamma}$. This is true since a bias of at most Γ_k implies bias at most Γ'_k for $\Gamma_k \leq \Gamma'_k$. Let $\overline{\mathbf{\Gamma}} \in \mathfrak{J}$ be such that $H^{k|K}_{0,\overline{\mathbf{\Gamma}}}$ is true and if $\mathbf{\Gamma} \in \mathfrak{J}$ and $H^{k|K}_{0,\mathbf{\Gamma}}$ is true then $\mathbf{\Gamma} \succ \overline{\mathbf{\Gamma}}$. $\overline{\mathbf{\Gamma}}$ might be empty, in which case there is nothing to prove.

Next, we note that $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}$ is increasing in $\mathbf{\Gamma}$; $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}} \leq P^{k|K}_{\mathbf{a},\mathbf{\Gamma}'}$ for $\mathbf{\Gamma} \leq \mathbf{\Gamma}'$. A rejection of a true null hypothesis when the corresponding maximum p-value is less than α, implies $P^{k|K}_{\mathbf{a},\overline{\mathbf{\Gamma}}} < \alpha$. Thus, the probability of rejecting any true null hypothesis is upper bounded by $\Pr(P^{k|K}_{\mathbf{a},\overline{\mathbf{\Gamma}}} < \alpha)$, which is at most α by Proposition 2. ■

The following corollary to the theorem considers a sensitivity analysis with the same bias parameter for all the factors. The proof of the following two corollaries are given in the appendix.

Corollary 1 *Assume the same conditions as in Theorem 3, except let* $\mathfrak{J} = \{\mathbf{\Gamma} = \Gamma_l(1,\ldots,1) : 1 = \Gamma_1 < \Gamma_2 < \cdots < \Gamma_L\}$. *Fix* $\mathbf{a} \in \{0,1\}^K$ *and* k, $1 \leq k \leq K$. *Consider the testing procedure that rejects* $H^{k|K}_{0,\mathbf{\Gamma}}$ *for* $\mathbf{\Gamma} \in \mathfrak{J}$ *if* $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}} < \alpha$. *Then the probability of rejecting any true null hypothesis among the set of hypotheses* $\{H^{k|K}_{0,\mathbf{\Gamma}}; \mathbf{\Gamma} \in \mathfrak{J}\}$ *is at most* α.

This corollary is relevant to the analyses of the lead example revisited in §7, which considers $\Gamma_1 = 1, \Gamma_2 = 1.2, \ldots, \Gamma_{11} = 3, \Gamma_{12} = 4, \Gamma_{13} = 4.8$ and $\Gamma_{11} = 5$, see Table 3. The final corollary combines the situations of Proposition 3 and Theorem 3.

Corollary 2 *Assume that conditions (a)–(c) of Proposition 3 are satisfied and assume the structure of* $\mathfrak{J}$ *either as in Theorem 3 or as in Corollary 1. Fix* $\mathbf{a} \in \{0,1\}^K$. *Consider the procedure that rejects* $H^{k|K}_{0,\mathbf{\Gamma}}$ *for* $\mathbf{\Gamma} \in \mathfrak{J}$ *if* $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}} < \alpha$. *Then the probability of rejecting any true null hypothesis among* $\{H^{k|K}_{0,\mathbf{\Gamma}}; 1 \leq k \leq K, \mathbf{\Gamma} \in \mathfrak{J}\}$ *is at most* α.

After rejecting a partial conjunction hypothesis it could be of interest to test the individual hypotheses, asking, if at least k of the K hypotheses are false, i.e., if $H^{k|K}_0$ is rejected, which of the individual hypotheses are false? The following proposition states that, in the multiparameter sensitivity analysis, the individual hypotheses can be tested, after rejecting the partial conjunction hypothesis $H^{k|K}_{0,\mathbf{\Gamma}}$, with a correction factor $(K-k)$.

Proposition 5 *Consider the setting of Theorem 3. Consider the testing procedure that rejects* $H^{k|K}_{0,\mathbf{\Gamma}}$ *for* $\mathbf{\Gamma} \in \mathfrak{J}$ *if* $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}} < \alpha$*; and when* $H^{k|K}_{0,\mathbf{\Gamma}}$ *is rejected, for* $1 \leq t \leq K$ *the procedure further rejects the hypothesis* $H_{0,t}$

when the bias is at most Γ_t *if* $\overline{P}_{t,\mathbf{a}_{t-1},\Gamma_t} < \alpha/(K-k)$*. Then the probability that this testing procedure rejects any true null hypothesis is at most* α*.*

Proof. By Theorem 3, the probability that the procedure rejects any true hypothesis in $\{H_{0,\mathbf{\Gamma}}^{k|K}; \mathbf{\Gamma} \in \mathfrak{J}\}$ is at most α. Suppose that a hypothesis $H_{0,t}$ when the bias is at most Γ_t is falsely rejected. For this to happen the procedure must first reject the hypothesis $H_{0,\mathbf{\Gamma}}^{k|K}$. There are two possibilities. First, $H_{0,\mathbf{\Gamma}}^{k|K}$ is true. In which case the probability of the false rejection is controlled by Theorem 3.

Otherwise, $H_{0,\mathbf{\Gamma}}^{k|K}$ is false. Then, by definition, at most $K-k$ individual hypotheses are true and the tth hypothesis is one of them. This also implies that the minimum bias in factor t is at least Γ_t. Suppose $\bar{\Gamma}_t$ is the true bias in factor t. Then $\bar{\Gamma}_t < \Gamma_t$. Thus, the rejection due to $\overline{P}_{t,\mathbf{a}_{t-1},\Gamma_t} < \alpha/(K-k)$, implies $\overline{P}_{t,\mathbf{a}_{t-1},\bar{\Gamma}_t} < \alpha/(K-k)$, as the individual sensitivity analysis p-values are increasing in the sensitivity parameters. Thus, the probability of rejecting any true null hypothesis of the K individual hypotheses is bounded by the probability of rejecting at least one of at most $K-k$ true null hypotheses $H_{0,t}$ and a bias of at most $\bar{\Gamma}_t$. This probability is less than the sum of the probability of rejecting each of them, which is less than $(K-k) \times \alpha/(K-k) = \alpha$. ■

Therefore, in our lead example, where $K = 5$, under the setting of Corollary 2, if we have evidence for at least 3 of 5 pieces of the elaborate theory, we can test the 5 individual pieces of the theory by comparing the separate sensitivity analyses p-values to $\alpha/2$. By comparison, a Bonferroni correction would have compared the separate sensitivity analyses p-values to $\alpha/5$.

6. Comparison of combining methods

6.1. Settings under which power of sensitivity analysis is judged

In a sensitivity analysis to unmeasured confounding, there are some situations in which it is clear what we would like a procedure to do and some situations in which the desired answer is unclear. An example of one of the latter situations is when there is large bias from unmeasured confounding and a treatment effect — we are nearly assured to reject the null for moderate values of the sensitivity parameter, but, such a rejection decision is not unambiguously sought after as we would also have rejected the null with moderate bias when the null is indeed true. One of the former situations, in which we are clear about the desired answer of the sensitivity analysis, is when there is a treatment effect and no bias from unmeasured confounding.

In this situation, a sensitivity analysis with a chosen value of the sensitivity parameter checks whether we are still able to reject the null, allowing for the level of bias given by the sensitivity parameter. It is desired then that a method is not fooled by moderate values of the sensitivity parameter and rejects the null. This situation has been called the "favorable situation" and is the situation under which power of sensitivity analysis has been evaluated (Rosenbaum, 2010; Hansen et al., 2014).

One might wonder why we evaluate the power of a sensitivity analysis under a setting in which there is actually no bias from unmeasured confounding when the sensitivity analysis is worried about bias. The reason is that, in most observational studies, we are worried about bias and cannot know that there is no bias, but we would like to have high power to say that we have evidence for a treatment effect that is insensitive to moderate bias if in fact there is a treatment effect and no bias.

In §6.2, we will analyze the asymptotics of power of sensitivity analysis when the sample size goes to infinity. There we provide a characterization of the asymptotically optimal choice of combining method, and find asymptotically optimal combining methods. In §6.3, we compare the combining methods in their power of sensitivity analysis using a simulation study. Since in practice we only have a finite sample, looking at the power of sensitivity analysis for finite samples might give us more guidance about the choice of method for analysis.

6.2. Asymptotically optimal tests

When there is a treatment effect and no unmeasured confounding, a method is preferred that can withstand larger bias in sensitivity analysis. When the sample size goes to infinity, this threshold of the sensitivity parameter is quantified as the design sensitivity of the method (Rosenbaum, 2004; Rosenbaum, 2010; Hsu et al., 2013; Hansen et al., 2014; Zhao, 2018). However, for partial conjunction testing from K evidence factors, design sensitivity for the various combining methods is a crude criterion of comparison. As we will see in Proposition 6 below, most combining methods have the same design sensitivity. Instead, we look at the rate of rejection for the combining methods in their sensitivity analysis when there is treatment effect and no unmeasured confounding. This rate of rejection is the Bahadur slope of a sensitivity analysis (Rosenbaum, 2015). The ratio of the slopes of two competing methods of analysis is called the Bahadur efficiency of sensitivity analysis. A method with larger slope needs a smaller sample size to make the desired decision with high probability (Bahadur, 1967; Rosenbaum, 2015; Ertefaie et al., 2018). In the following, we show that Fisher's method and the trun-

cated product method are optimal in this regard. Put differently, Fisher's method (Fisher, 1932) and the truncated product method (Zaykin, 2002) have Bahadur efficiency of sensitivity analysis one, relative to each other, and have efficiency at least one, relative to any other combining method.

We first introduce some notation to facilitate the discussion. Recall, the partial conjunction p-values are defined for a set of functions $(g_1, \ldots, g_K)$ where $g_k : [0,1]^{K-k+1} \to [0,1]$, $k = 1, \ldots, K$, as

$$P^{k|K}_{\mathbf{a},\mathbf{\Gamma}} = g_k(\overline{P}_{(k)\mathbf{a},\mathbf{\Gamma}}, \ldots, \overline{P}_{(K)\mathbf{a},\mathbf{\Gamma}}).$$

Here, $\overline{P}_{(1)\mathbf{a},\mathbf{\Gamma}} \leq \cdots \leq \overline{P}_{(K)\mathbf{a},\mathbf{\Gamma}}$ are the ordered values of $\overline{P}_{1,\mathbf{a}_0,\Gamma_1}, \ldots, \overline{P}_{K,\mathbf{a}_{K-1},\Gamma_K}$. Now we emphasize the choice of the combining functions by denoting $\mathfrak{g} = (g_1, \ldots, g_K)$ and using $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{g})$ to denote the above quantity. We use the notation $\mathfrak{ef} = (f_1, \ldots, f_K)$ to denote Fisher's combining functions. That is, the kth function in $\mathfrak{ef}$ is $f_k(x_k, \ldots, x_K) = \Pr(\chi^2_{2(K-k+1)} > -2\sum_{j=k}^{K} \log x_j)$. The optimality statement made in this section is an asymptotic statement. We must think of $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k}$ as function of I, the number of pairs. Consequently, $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{g})$ is also a function of I. These dependencies will not be made explicit below. The asymptotic here is with K fixed and I going to infinity.

Consider the situation where there is an effect, i.e., some of the K hypotheses $H_{0,k}$ are false. Suppose, there is no unmeasured confounding. We noted that the desired result of a sensitivity analysis, in this situation, is to be able to reject the null. Suppose $H_{0,k}$ is false. The maximum p-value for the kth factor is $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k}$. For any sample size, as $\Gamma_k \to \infty$ this maximum p-value $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \to 1$, a formal statement for the known fact that any treatment effect, however large, can be explained by large enough bias. The design sensitivity for this factor is the bias level $\widetilde{\Gamma}_k$ such that $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \to 0$ for $\Gamma_k < \widetilde{\Gamma}_k$ and $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \to 1$ for $\Gamma_k > \widetilde{\Gamma}_k$; the limit here is with $I \to \infty$. For example, in the lead study, each test for the 5 pieces of the elaborate theory has a design sensitivity. When a piece of the theory is true, then with sufficient sample the test would provide evidence for it as long as, and only when, the bias level is less than the design sensitivity of the test.

Now we look at the sensitivity analysis for the partial conjunctions of these evidence factors. The following proposition studies the design sensitivity of this multi-parameter sensitivity analysis, and concludes that most methods are indistinguishable in this regard.

Proposition 6 *Take any combining method* $\mathfrak{g}$. *Suppose,* $g_k(0, \ldots) = 0$ *and* $g_k(1, \ldots, 1) = 1$ *and* g_k *is continuous at* $\{0,1\}^{K-k+1}$. *With the sensitivity parameter* $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_K)$ *for the partial conjunction testing, we have*

$P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{g}) \to 1$ *if* $\widetilde{\Gamma}_l < \Gamma_l$ *for* $K-k+1$ *many* Γ_l. *Also,* $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{g}) \to 0$ *if* $\Gamma_l < \widetilde{\Gamma}_l$ *for at least* k *many* Γ_l *and* $\Gamma_l \neq \widetilde{\Gamma}_l$ *for all* l.

The following theorem says that, in the class of functions for $\mathfrak{g}$ considered in §5, Fisher's method, $\mathfrak{ef}$, has the optimal Bahadur slope.

Assumption: A sequence of numbers $c(I)$ satisfies $c(I) \to \infty$ as $I \to \infty$. As I increases to infinity, $c(I)^{-1} \log \overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \to -r_k(\Gamma_k)$ almost surely, where $r_k(\Gamma_k) \in [0,\infty]$, for $k = 1,\ldots,K$. We call $r_k(\Gamma_k)$ the slope of test k at Γ_k.

Theorem 4 *Consider any set of* K *combining functions* $\mathfrak{g} = (g_1,\ldots,g_K)$ *such that each* g_k *is coordinatewise nondecreasing and satisfies* $\Pr(g_k(U_k,\ldots,U_K) \leq \alpha) \leq \alpha$, *for any* $\alpha \in [0,1]$, *for i.i.d. uniform(0,1) random variables* $U_1,\ldots,U_K$; $k = 1,\ldots,K$. *We have, for Fisher's combining method* $\mathfrak{ef} = (f_1,\ldots,f_K)$,

$$\lim_{I\to\infty} c(I)^{-1} \log P^{k'|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{ef}) \leq \liminf_{I\to\infty} c(I)^{-1} \log P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{g}) \qquad \text{for } k' \leq k$$

almost surely for $k, k' = 1,\ldots,K$.

The assumption talks about the Bahadur slope of sensitivity analysis for the individual factors. Rosenbaum (2015) provides a detailed discussion on the existence and calculation of the limit. The limit depends on the choice of the test statistic, the joint distribution of the potential outcomes for the units, and the distribution of the treatment assignment. The above assumption and the theorem while general also allow us to consolidate several important implications.

Following Proposition 6, our interest is in the case when we are able to reject the null in the sensitivity analysis, when in truth there is an effect. This is the case for a sensitivity parameter $\mathbf{\Gamma}$ with some of the bias levels less than the design sensitivity. Let $\widetilde{k}$ be the number of Γ_l with $\Gamma_l < \widetilde{\Gamma}_l$. Any method in Proposition 6 will reject $H_0^{k|K}$ whenever $k \leq \widetilde{k}$, as the sample size goes to infinity. The rate of rejection is used in Theorem 4 to tell the combining methods apart. The following Proposition finds the slope of Fisher's method. This slope is the same as that of the truncated product method with a truncation level $\varkappa$, and is at least as large as any other method that satisfies the conditions of Theorem 4.

Proposition 7 *Suppose there is no unmeasured confounding and* H_0 *is false. Let the design sensitivity of the test* k *be* $\widetilde{\Gamma}_k$. *Consider a sensitivity analysis with sensitivity parameter* $\mathbf{\Gamma}$ *such that* $\Gamma_k \neq \widetilde{\Gamma}_k$ *for all* k. *Let* $\widetilde{k}$ *be the number of* Γ_l *with* $\Gamma_l < \widetilde{\Gamma}_l$. *Finally, let* $r_{(1)\mathbf{\Gamma}} \leq \cdots \leq r_{(K)\mathbf{\Gamma}}$ *are ordered values of* $r_1(\Gamma_1),\ldots,r_K(\Gamma_K)$. *We have* $\lim_{I\to\infty} c(I)^{-1} \log P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{ef}) =$

$-\mathbf{1}(k \leq \widetilde{k}) \sum_{K-\widetilde{k}+1}^{K-k+1} r_{(l)\mathbf{\Gamma}}$. *The truncated product method with* $\varkappa \in (0, 1]$ *has the same slope as Fisher's method.*

6.3. Simulation study: Finite sample power of sensitivity analysis

Section 5 discussed various choices of the function g_k, which is used to define $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}$. In this section we compare these combining methods in their power of sensitivity analysis in finite samples using a simulation study.

In the simulation setting we set $I = 150$ and $K = 5$. Treatment k has an additive effect β_k and we assume a standard normal variate for the base response in the absence of any treatment. Thus, when a unit has been assigned treatment $(z_1, \ldots, z_K)$, a binary vector of length K, the response of that unit is $\sum_{k=1}^{K} z_k\beta_k + N(0, 1)$. We simulate a treatment assignment which is random, thus within each pair, each unit has probability $1/2$ of getting the first treatment. Further, the treatments are simulated to be independent of each other in a way that $Z^{(k)}_{ij} \overset{i.i.d}{\sim} Bernoulli(0.6)$ for $2 \leq k \leq 4$ and $Z^{(5)}_{ij} \sim Bernoulli(0.5)$. In the appendix we present more simulation results exploring other data generating processes, varying I, using different number of treatments, using correlated treatment assignments, and by varying the model of the response.

In the power of sensitivity analysis, we look at the simulated power of of rejecting $H^{k|K}_0$ for the various methods when we assume various Γ values for bias. A method is less sensitive if, in the presence of a treatment effect, it maintains power to detect that treatment effect at higher values of Γ (Rosenbaum, 2004). We take $\mathbf{a} = (1, 1, 1, 1, 1)$ as in the §2. The basic tests use Wilcoxon's paired sample and two sample statistics. These simulation results are presented in Table 1, where each sampling situation was replicated 15,000 times, so that a binomial proportion has a standard error less than $\sqrt{0.25/15000} \approx 0.004$. The four methods compared in the simulation are Holm-Bonferroni method (henceforth Holm's), Simes' method, the modified additive p-value method (henceforth SumP) and the truncated product method. Holm's method ignores the near independence of the separate analyses established in Theorem 2. For Holm's method $g_k(x_k, \ldots, x_K) = (K - k + 1)x_k$ (Holm, 1979). Simes' method and the SumP method satisfy the desired conditions of Proposition 3, Holm's method does not. For the truncated product method we consider the familiar level of truncation $\varkappa = 0.2$. This method was further modified to redefine $P^{k|K}_{\mathbf{a},\mathbf{\Gamma}} = \max\{P^{1|K}_{\mathbf{a},\mathbf{\Gamma}}, \ldots, P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}\}$ for $k = 1, \ldots, K$, so that it provided monotone p-values, required in Proposition 3. In Table 1, a simulated power

of 0 is replaced by a blank cell for ease of viewing.

Table 1 does not show the results of a näive method that only counts the number of hypotheses rejected when each sensitivity analyses is compared to level 0.05. Because, this method does not provide control of the type-I error for the testing problem. In the null case, scenario 1, in our simulations, the probability of rejecting $H_0^{1|5}$ is 0.224, while the expected level is 0.05. Holm's method, whose simulated power is reported in the table, is the modification of the näive method that controls the type-I error rate.

There are at least two ways of reading Table 1. First, we look at each of the methods individually and compare the various scenarios of treatment effect. Note that, the power for each of the methods decrease as we read the table from right to left, increasing the value of k, and top to bottom in each scenario, increasing the value of Γ. The null case of no treatment effect, scenario 1, is a check that the analysis is performed at level of significance 0.05 and the methods control the type 1 error. Across the scenarios, moving from the null scenario to the scenario where each treatment has an effect of size 0.25 (scenario 3), the simulated power increases for each of the methods. The power of rejecting at least 3 basic hypotheses out of 5, $H_0^{3|5}$, for $\Gamma = 1$, is 9% for SumP method in Scenario 2 and 32% in Scenario 3. The corresponding numbers are 5% and 10% for the Simes' method, and 7% and 23% for the truncated product method.

Consider a second perspective to Table 1. We compare the methods within the various scenarios. The power of the SumP method is much smaller in rejecting $H_0^{1|5}$ ($k = 1$) compared to the other methods. The power, in scenario 2 with $\Gamma = 1$, is 57% for SumP compared to 99% for Holm's, Simes', and the truncated product method. Also, in terms of the maximum bias level of sensitivity analysis a method can tolerate, (which, one can read by looking at the level of bias where the numbers in the column first vanishes) Holm's and Simes' method are less sensitive when $k = 1$ for both scenario 2 and 3. The story is somewhat reversed for larger k. For example, consider $k = 3$ or $H_0^{3|5}$ in scenario 3. The simulated power for $\Gamma = 1$ is highest for SumP (32%) and lowest for Holm's method (8%) and second lowest for Simes' method (10%); for the truncated product it is 23%. Further, SumP is less sensitive (sensitive at $\Gamma = 2$) compared to Simes', and Holm's method (sensitive at $\Gamma = 1.6$) and the truncated product method (sensitive at $\Gamma = 1.8$).

To summarize, no one method is victorious. But it seems Simes' or Holm's method is a poor choice as they lose their power fast going from right to left of the table. Holm's method essentially looks at the individual p-values and does not pool them, thus it often misses that there is evidence for some

TABLE 1

Simulation results for the power of sensitivity analysis evaluated at level 0.05. Numbers are out of 100. A cell value is the percentage of times the decision that at least k many $H_{0,l}$s are false is made, with $\Gamma_1 = \cdots = \Gamma_5 =: \Gamma$, out of 15000 simulations. Empty cells represent the value 0. tP = truncated product method with truncation level $\varkappa = 0.20$; sP = the modified additive p-value method in Proposition 4; Si = Simes' method; HB = Holm-Bonferroni method.

$k \rightarrow$	5				4				3				2				1			
$\Gamma \downarrow$	tP	sP	Si	HB	tP	sP	Si	HB	tP	sP	Si	HB	tP	sP	Si	HB	tP	sP	Si	HB
Scenario 1: (null case) $\beta_1 = \cdots = \beta_5 = 0$																				
1	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	0	5	5	5	5
Scenario 2: $\beta_1 = \beta_2 = \beta_3 = 0.25, \beta_4 = \beta_5 = 0$																				
1					1	2			7	9	5	5	40	28	33	32	99	57	99	99
1.2									2	3	1	1	18	12	16	16	90	34	94	93
1.4										1			7	4	7	7	66	16	78	77
1.6													2	2	2	2	38	7	55	55
1.8													1	1	1	1	18	3	35	34
2																	7	1	19	19
2.2																	3		10	10
2.4																	1		4	4
2.6																			2	2
2.8																			1	1
3																				
Scenario 3: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0.25$																				
1					6	10	2	1	23	32	10	8	59	61	42	39	100	84	100	100
1.2					1	3			8	15	3	3	34	37	23	21	100	64	100	100
1.4						1			3	6	1	1	17	19	12	11	99	43	100	100
1.6									1	2			8	9	6	6	95	26	99	99
1.8										1			4	4	3	3	83	14	96	95
2													2	1	2	2	65	7	89	89
2.2													1	1	1	1	46	3	78	78
2.4															1	1	29	1	65	64
2.6																	17	1	50	50
2.8																	9		37	36
3																	5		26	25
3.6																	1		8	7
4																			3	3

fraction of the nulls not being true when each test does not have sufficient power. While SumP has a much smaller power in providing evidence that at least one of the nulls is false, it retains a lot of its power when looking for more pieces of evidence (going right to left). The truncated product method seems be a fair compromise based on these simulations.

TABLE 2
The p-values, under the assumption of no unmeasured confounding, for testing the hypothesis that at least k many $H_{0,l}$s are false in the lead absorption study. $K = 5$ and $\Gamma_1 = \cdots = \Gamma_5 = 1$. SumP = the modified additive p-value method in Proposition 4.

	method			
k	Simes'	SumP	Fisher's	Truncated Product ($\varkappa = 0.20$)
5	0.420036	0.420036	0.420036	1
4	0.191846	0.133107	0.169691	0.193477
3	0.028322	0.024172	0.015168	0.017172
2	0.015242	0.003268	0.000739	0.000795
1	0.000348	0.000346	$1.41 \cdot 10^{-6}$	$1.57 \cdot 10^{-6}$

7. Revisiting the lead absorption study

The p-values for the five tests were reported in §2 for the causal hypothesis that occupational exposure to lead increases the lead level in the blood of the children. If there is no bias due to unmeasured confounding, i.e., assuming $\Gamma_1 = \cdots = \Gamma_5 = 1$, these p-values are $P_1 = 2.69 \cdot 10^{-5}$, $P_2 = 3.81 \cdot 10^{-3}$, $P_3 = 9.59 \cdot 10^{-2}$, $P_4 = 9.44 \cdot 10^{-3}$, and $P_5 = 0.42$. The p-values for the tests for partial conjunction of the hypotheses are given in Table 2. This table reports the results from four methods of pooling evidence. Qualitatively, the results from the four methods are similar. At $\alpha = 0.05$, we have evidence for rejecting at least 3 out of 5 basic nulls. The p-values from Fisher's method and truncated product method are much smaller when compared to the other methods.

How sensitive are these tests to unmeasured confounding? The maximum p-values for the five tests are presented at the top half of Table 3. At significance level 0.05, of the five tests, the first, second, and the fourth test rejects the corresponding hypotheses, assuming no bias from unmeasured confounding. These tests become sensitive at bias levels $\Gamma_1 = 4.8$, $\Gamma_2 = 2.8$, and $\Gamma_4 = 3$, respectively. But, this is an incorrect interpretation of the results. The type-I error is at most 0.05 in each column. But, across the rows the type-I error is not controlled in this top half of Table 3. If we control for the type-I error using Bonferroni correction, we would compare the maximum p-values to $0.05/5 = 0.01$. Thus, in the top half of the table, maximum p-values less than 0.01 are highlighted in bold. The first test becomes sensitive at $\Gamma_1 = 2.6$, the second test at $\Gamma_2 = 1.4$ and the fourth test is sensitive even at $\Gamma_4 = 1.2$.

The bottom half of Table 3 presents a sensitivity analysis for the partial conjunctions of the tests. By Corollary 2, this part of the table provides

TABLE 3

Evidence factors analysis of the lead absorption study. (1) The first half of the table: Maximum p-values corresponding to the five tests with $\Gamma_k = \Gamma$, $1 \le k \le 5$. We dropped the subscript $\mathbf{a}_{k-1}$ from $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma}$ used in Section 3.3–Section 6. (2) The second half of the table: Maximum p-values for testing at least k of $H_{0,l}$s are false when the bias is at most $\Gamma_1 = \cdots = \Gamma_5 = \Gamma$ (using the truncated product method; truncation level $\varkappa = 0.20$). The maximum p-values less than 0.05 in the lower half, and less than $0.05/5 = 0.01$ in the upper half are highlighted in bold.

$\Gamma\downarrow$	$\overline{P}_{5,\Gamma}$	$\overline{P}_{4,\Gamma}$	$\overline{P}_{3,\Gamma}$	$\overline{P}_{2,\Gamma}$	$\overline{P}_{1,\Gamma}$
1	0.420036	**0.009441**	0.095923	**0.00381**	**0.00007**
1.2	0.470253	0.013512	0.128619	**0.006773**	**0.000263**
1.4	0.512934	0.017814	0.161157	0.010557	**0.000688**
1.6	0.549884	0.022219	0.192914	0.015089	**0.001425**
1.8	0.582428	0.02672	0.223553	0.020268	**0.002525**
2	0.611224	0.031257	0.252909	0.025994	**0.004007**
2.2	0.636902	0.035769	0.280914	0.032177	**0.005867**
2.4	0.659949	0.040228	0.307565	0.038738	**0.008085**
2.6	0.680756	0.044615	0.332889	0.045607	0.010632
2.8	0.699635	0.048916	0.356935	0.052721	0.013472
3	0.716841	0.053123	0.379764	0.060029	0.016569
4	0.784073	0.072632	0.477894	0.098608	0.034756
4.8	0.822295	0.08707	0.541509	0.130282	0.051015
5	0.830333	0.090589	0.555832	0.138152	0.055166
$\Gamma\downarrow$	$P_\Gamma^{5\|5}$	$P_\Gamma^{4\|5}$	$P_\Gamma^{3\|5}$	$P_\Gamma^{2\|5}$	$P_\Gamma^{1\|5}$
1	1	0.193477	**0.017172**	**0.000795**	**0.000002**
1.2	1	0.24579	**0.027005**	**0.001965**	**0.000012**
1.4	1	0.297852	**0.037873**	**0.003864**	**0.000052**
1.6	1	0.348663	**0.049288**	**0.006544**	**0.00016**
1.8	1	1	0.149304	**0.026378**	**0.001114**
2	1	1	0.161532	**0.033879**	**0.002024**
2.2	1	1	0.17212	**0.041805**	**0.003305**
2.4	1	1	0.181238	0.050012	**0.004979**
2.6	1	1	0.191565	0.05838	**0.007052**
2.8	1	1	0.205224	0.066817	**0.009511**
3	1	1	0.219255	0.075254	**0.012336**
4	1	1	0.293328	0.116672	**0.030932**
4.8	1	1	0.354142	0.148886	**0.049496**
5	1	1	0.369251	0.156729	0.054454

an adaptive analysis, in the sense that the total type-I error is at most 0.05. Learning from the results of the simulation study in §6 we chose the truncated product method with truncation level 0.20 in computing the partial conjunction p-values. When the bias is at most $\Gamma = 1.6$ we have evidence to reject at least 3 of the 5 basic nulls. When the bias is at most $\Gamma = 2$ we no longer have evidence to reject 3, but the evidence allows us to reject 2 of the 5 basic nulls.

We can also look at the individual tests after observing that at $\Gamma = 1.6$ we reject at least 3 out of the 5 nulls, by comparing them to $0.05/2 = 0.025$, see Proposition 5. Three basic nulls, the first, second and fourth, are rejected by this procedure with maximum p-values 0.001425, 0.015089 and 0.022219 respectively.

8. Conclusion

Study of a causal hypothesis is enhanced when directed tests are considered for the various predictions of the hypothesis. Of course, these testable predictions of a causal hypothesis would be based on acknowledged theories at the time when the causal hypothesis is being investigated. Inherent to these predictions are requirements of simplicity and falsifiability.

On the other spectrum of etiology, a statistical analysis of a causal or etiologic hypothesis should focus on comprehensive reports that help explicate the step from an observed data to corroboration of the hypothesis. With this aim, this paper presents a method of analysis of an elaborate theory of predictions of a causal hypothesis. We consider such elaborate theories whose falsifiable statements can be set up as alternative hypotheses in statistical hypothesis testing problems. An etiologic hypothesis can still be false because some other prediction of the hypothesis is not true. But the focus of this paper has been to assess the extent to which the observed data supports the predictions in the elaborate theory. Our analysis suggests decomposing the tests of the elaborate theory into nearly independent factors. Partial conjunctions of these tests tell us about fractions of the elaborate theory. As the tests might themselves be biased by unmeasured confounding, we also consider a multi-parameter sensitivity analysis. We are thus able to quantify the bias levels at which the observed data supports a certain fraction of the elaborate theory. When the tools of this analysis are appropriately chosen, the overall type-I error of this analysis is controlled without having to pay a price for having considered multiple tests, thus, without losing any power.

Appendix

Proof of Theorem 2. Let $\mathcal{K} = \{k_1, \ldots, k_{|\mathcal{K}|}\} \subseteq \{1, \ldots, K\}$ and $U_1, \ldots, U_{|\mathcal{K}|}$ be $|\mathcal{K}|$ i.i.d. random variables uniform on $[0, 1]$. Since $\overline{P}_{k_1, \mathbf{a}_{k_1-1}, \Gamma_{k_1}}$ is the maximum of $\overline{P}_{k_1, \mathbf{a}_{k_1-1}}$ over the unmeasured confounders u_{ijk_1}'s. For $\alpha_1 \in$

$[0,1]$ we have

$$\begin{aligned}
&\Pr(\overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}} \le \alpha_1 \mid \mathcal{F}, H_{0,k_1}) \\
&\le \Pr(P_{k_1,\mathbf{a}_{k_1-1}} \le \alpha_1 \mid \mathcal{F}, H_{0,k_1}) \\
&\le E[\Pr(P_{k_1,\mathbf{a}_{k_1-1}} \le \alpha_1 \mid \mathbf{Z}_{k_1-1}, \sum_{ij\in\mathcal{I}_{k_1-1}(\mathbf{a}_{k_1-1})} Z_{ij}^{(k_1)}, \mathcal{F}, H_{0,k_1})] \le E[\alpha_1] = \Pr(U_1 \le \alpha_1).
\end{aligned}$$

The expectation in the previous calculation is over the joint distribution of $\mathbf{Z}_{k_1-1}, \sum_{ij\in\mathcal{I}_{k_1-1}(\mathbf{a}_{k_1-1})} Z_{ij}^{(k_1)}$ conditional on $\mathcal{F}, H_{0,k_1}$. We borrow the notation of Shaked and Shanthikumar (2007). Then $U_1 \le_{st} \overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}}$.

Now let $2 \le l \le |\mathcal{K}|$. Note that for any k_l the maximum p-value $\overline{P}_{k_l,\mathbf{a}_{k_l-1},\Gamma_{k_l}}$ is a function of $\mathcal{Z}_l$ and $\mathcal{F}$. Hence, for $\alpha_l \in [0,1]$,

$$\begin{aligned}
&\Pr(\overline{P}_{k_l,\mathbf{a}_{k_l-1},\Gamma_{k_l}} \le \alpha_l \mid \overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}}, \ldots, \overline{P}_{k_{l-1},\mathbf{a}_{k_{l-1}-1},\Gamma_{k_{l-1}}}, \mathcal{F}, H_{0,k_l}) \\
&\le \Pr(P_{k_l,\mathbf{a}_{k_l-1},\Gamma_{k_l}} \le \alpha_l \mid \overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}}, \ldots, \overline{P}_{k_{l-1},\mathbf{a}_{k_{l-1}-1},\Gamma_{k_{l-1}}}, \mathcal{F}, H_{0,k_l}) \\
&\le E\big[\Pr(P_{k_l,\mathbf{a}_{k_l-1},\Gamma_{k_l}} \le \alpha_l \mid \mathbf{Z}_{l-1}, \sum_{ij\in\mathcal{I}_{k_l-1}(\mathbf{a}_{k_l-1})} Z_{ij}^{(k_l)}, \overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}}, \\
&\qquad \ldots, \overline{P}_{k_{l-1},\mathbf{a}_{k_{l-1}-1},\Gamma_{k_{l-1}}}, \mathcal{F}, H_{0,k_l})\big] \\
&\le E\big[\Pr(P_{k_l,\mathbf{a}_{k_l-1},\Gamma_{k_l}} \le \alpha_l \mid \mathbf{Z}_{l-1}, \sum_{ij\in\mathcal{I}_{k_l-1}(\mathbf{a}_{k_l-1})} Z_{ij}^{(k_l)}, \mathcal{F}, H_{0,k_l}) \,\big| \\
&\qquad \overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}}, \ldots, \overline{P}_{k_{l-1},\mathbf{a}_{k_{l-1}-1},\Gamma_{k_{l-1}}}, \mathcal{F}, H_{0,k_l}\big] \\
&\le E[\alpha_l \mid \overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}}, \ldots, \overline{P}_{k_{l-1},\mathbf{a}_{k_{l-1}-1},\Gamma_{k_{l-1}}}, \mathcal{F}, H_{0,k_l}] \\
&\le \alpha_l = \Pr(U_l \le \alpha_l).
\end{aligned}$$

Thus under $\cap_{t\in\mathcal{K}} H_{0,t}$ and conditional on $\mathcal{F}$,

$$U_l \le_{st} [\overline{P}_{k_l,\mathbf{a}_{k_l-1},\Gamma_{k_l}} \le \alpha_l \mid \overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}}, \ldots, \overline{P}_{k_{l-1},\mathbf{a}_{k_{l-1}-1},\Gamma_{k_{l-1}}}]$$

for all $2 \le l \le |\mathcal{K}|$.

Also, $(U_1, \ldots, U_{|\mathcal{K}|})$ is a conditionally increasing in sequence (CIS) (see, eq 6.B.11 of Shaked and Shanthikumar (2007)). Thus, by Theorem 6.B.4 of Shaked and Shanthikumar (2007) under $\cap_{t\in\mathcal{K}} H_{0,t}$ and conditional on $\mathcal{F}$, $(U_1, \ldots, U_{|\mathcal{K}|}) \le_{st} (\overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}}, \ldots, \overline{P}_{k_{|\mathcal{K}|},\mathbf{a}_{k_{|\mathcal{K}|}-1},\Gamma_{k_{|\mathcal{K}|}}})$.

Let $U \subset \mathbb{R}^{|\mathcal{K}|}$ be called an upper set if, $\mathbf{x} \in U$ and $\mathbf{y} \gtrsim \mathbf{x}$ implies $\mathbf{y} \in U$. Then, we have for any upper set of the $|\mathcal{K}|$ dimensional euclidean space $\Pr((U_1, \ldots, U_{|\mathcal{K}|}) \in U) \le \Pr((\overline{P}_{k_1,\mathbf{a}_{k_1-1},\Gamma_{k_1}}, \ldots, \overline{P}_{k_{|\mathcal{K}|},\mathbf{a}_{k_{|\mathcal{K}|}-1},\Gamma_{k_{|\mathcal{K}|}}}) \in U)$.

Now to complete the proof set $U = \{(x_1, \ldots, x_{|\mathcal{K}|}) : f_{\mathcal{K}}(x_1, \ldots, x_{|\mathcal{K}|}) > x\}$ and note that U is an upper set since $f_{\mathcal{K}}$ is coordinatewise nondecreasing.

Proof of Proposition 4. 1. Consider first the 'minimum p-value' method. Condition (a) is obviously true. Next, note that $\Pr(\min\{U_k, \ldots, U_K\} \leq p) = 1-(1-p)^{K-k+1}$. Thus condition (b) is satisfied. Since, $\Pr(g_k(U_k, \ldots, U_K) \leq \alpha) = \Pr(\min\{U_k, \ldots, U_K\} \leq 1-(1-\alpha)^{1/(K-k+1)}) = 1-(1-(1-(1-\alpha)^{1/(K-k+1)}))^{(K-k+1)} = \alpha$. Finally, to check condition (c) fix $x_k \leq x_{k+1} \leq \cdots \leq x_K$. To check $g_k(x_k, \ldots, x_K) \leq g_{k+1}(x_{k+1}, \ldots, x_K)$, it is enough to show that $(1-x_k)^{K-k} - (1-x_{k+1})^{K-k+1} \geq 0$. This is true since, $(1-x_k)^{K-k} - (1-x_{k+1})^{K-k+1} \geq (1-x_{k+1})^{K-k} - (1-x_{k+1})^{K-k+1} = (1-x_{k+1})^{K-k} x_{k+1} \geq 0$.

2. Proofs of condition (a) and (b) are straightforward for Stouffer's method. To check condition (c) consider $x_k \leq x_{k+1} \leq \cdots \leq x_K$. Then, after some rearranging $g_k(x_k, \ldots, x_K) \leq g_{k+1}(x_{k+1}, \ldots, x_K)$ is equivalent to the inequality, $(\sqrt{(K-k+1)/(K-k)} - 1)(\Phi^{-1}(1-x_{k+1}) + \cdots + \Phi^{-1}(1-x_K)) \leq \Phi^{-1}(1-x_k)$. Since $x_k \leq \min\{x_{k+1}, \ldots, x_K\}$, it is enough to check that this condition holds with $x_k = 1-\Phi(\,(\Phi^{-1}(1-x_{k+1}) + \cdots + \Phi^{-1}(1-x_K))/(K-k)\,)$. Then the check reduces to checking $(\sqrt{(K-k+1)/(K-k)} - 1)(\Phi^{-1}(1-x_{k+1}) + \cdots + \Phi^{-1}(1-x_K)) \leq (\Phi^{-1}(1-x_{k+1}) + \cdots + \Phi^{-1}(1-x_K))/(K-k)$, or $(\sqrt{(K-k+1)/(K-k)} - 1) \leq 1/(K-k)$, or $\sqrt{1+1/(K-k)} \leq 1 + 1/(K-k)$; which is true.

3. Finally, consider the 'modified additive p-value' method. Condition (a) is obvious since g_k is an increasing function of $A_k = x_k + \cdots + x_K$. For condition (b) note from Edgington (1972), $\Pr(U_k + \cdots + U_K \leq x) \leq x^{K-k+1}/(K-k+1)!$. Let $F(x) := \Pr(U_k + \cdots + U_K \leq x)$. Then $F(x) \leq \min\{1, x^{K-k+1}/(K-k+1)!\} \leq \min\{1, x^{K-k+1}/(K-k+1)!\}^{1(x \leq c_k)}$. Thus, $\Pr(\min\{1, (U_k + \cdots + U_K)^{K-k+1}/(K-k+1)!\}^{1((U_k+\cdots+U_K) \leq c_k)} \leq \alpha) \leq \Pr(F(U_k + \cdots + U_K) \leq \alpha) \leq \alpha$.

For condition (c) fix $x_k \leq x_{k+1} \leq \cdots \leq x_K$. If $A_{k+1} = x_{k+1} + \cdots + x_K > c_{k+1}$, $g_{k+1}(x_{k+1}, \ldots, x_K) = 1$, thus the condition is satisfied. Suppose now $x_{k+1} + \cdots + x_K \leq c_{k+1}$. Clearly, $x_k \leq (x_{k+1} + \cdots + x_K)/(K-k) = A_{k+1}/(K-k)$; thus $g_k(x_k, \ldots, x_K) \leq g_k(A_{k+1}/(K-k), x_{k+1}, \ldots, x_K)$. Hence, it is enough to show that $g_k(A_{k+1}/(K-k), x_{k+1}, \ldots, x_K) \leq g_{k+1}(x_{k+1}, \ldots, x_K)$. Note that, $A_{k+1}/(K-k) + x_{k+1} + \cdots + x_K = A_{k+1}(K-k+1)/(K-k)$. Since, $A_{k+1} \leq c_{k+1}$, we get, $A_{k+1}(K-k+1)/(K-k) \leq c_k$. Hence, by simple reduction $g_k(A_{k+1}/(K-k), x_{k+1}, \ldots, x_K) \leq g_{k+1}(x_{k+1}, \ldots, x_K)$ is equivalent to $A_{k+1}^{K-k+1}(K-k+1)^{K-k}/(K-k)^{K-k+1} \leq A_{k+1}^{K-k}$; which simplifies to $A_{k+1} \leq (K-k)(1-1/(K-k+1))^{K-k} = c_{k+1}$. Thus proving condition (c).

Sketch of proof of Corollary 1. The proof is in line of the proof of Theorem 3 given in the main text. The main observation is that the thresholding level of the sensitivity parameter, $\bar{\mathbf{\Gamma}}$ exists even when $\mathfrak{J}$ is not a grid but a one dimensional hyper-plane $\mathfrak{J} = \{\mathbf{\Gamma} = \Gamma_l(1, \ldots, 1) : 1 = \Gamma_1 < \cdots < \Gamma_L\}$. Thus, probability of rejecting any of the true null among $\{H_{0,\mathbf{\Gamma}}^{k|K}; \mathbf{\Gamma} \in \mathfrak{J}\}$ is at most $\Pr(P_{\mathbf{a},\bar{\mathbf{\Gamma}}}^{k|K} \leq \alpha) \leq \alpha$.

Sketch of proof of Corollary 2. If there is no null among $\{H_0^{k|K}; 1 \leq k \leq K\}$ is true, there is nothing to prove. Otherwise, suppose $H_0^{t|K}$ is the first one in the list which is true. Recall that, under conditions (a)–(c) of Proposition 3, which is assumed in this corollary, for any $\mathbf{\Gamma}$ we have $P_{\mathbf{a},\mathbf{\Gamma}}^{1|K} \leq \cdots \leq P_{\mathbf{a},\mathbf{\Gamma}}^{K|K}$. Thus, rejection of any true null in $\{H_{0,\mathbf{\Gamma}}^{k|K}; \mathbf{\Gamma} \in \mathfrak{J}, 1 \leq k \leq K\}$ will mean that a true null in $\{H_{0,\mathbf{\Gamma}}^{t|K}; \mathbf{\Gamma} \in \mathfrak{J}\}$ is rejected. Define $\bar{\mathbf{\Gamma}}$ as in the proof of Theorem 3 or Corollary 1. Since $P_{\mathbf{a},\mathbf{\Gamma}}^{t|K}$ is nondecreasing in $\mathbf{\Gamma}$, rejecting any true null among $\{H_{0,\mathbf{\Gamma}}^{k|K}; \mathbf{\Gamma} \in \mathfrak{J}, 1 \leq k \leq K\}$ means rejecting $H_{0,\bar{\mathbf{\Gamma}}}^{t|K}$, which has probability at most α.

Proof of Proposition 6. Recall that $P_{\mathbf{a},\mathbf{\Gamma}}^{k|K}(\mathfrak{g}) = g_k(\overline{P}_{(k)\mathbf{a},\mathbf{\Gamma}}, \ldots, \overline{P}_{(K)\mathbf{a},\mathbf{\Gamma}})$. Consider the first case, $\Gamma_l > \widetilde{\Gamma}_l$ for at most k many l. It follows from the definition of design sensitivity that the largest $K - k + 1$ p-values converge to 1. Thus, $P_{\mathbf{a},\mathbf{\Gamma}}^{k|K}(\mathfrak{g}) \to g_k(1, \ldots, 1) = 1$. In the second case, $\Gamma_l < \widetilde{\Gamma}_l$ for k or more l's. By the definition of design sensitivity $\overline{P}_{(l)\mathbf{a},\mathbf{\Gamma}} \to 0$ for $l = 1, \ldots, k$ and the rest goes to 1. Thus $P_{\mathbf{a},\mathbf{\Gamma}}^{k|K}(\mathfrak{g}) \to g_k(0, \ldots) = 0$.

Proof of Theorem 4. By the assumption, $c(I)^{-1} \log \overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \to -r_k(\Gamma_k)$ almost surely for $k = 1, \ldots, K$. Let $r_{(1)\mathbf{\Gamma}} \leq \cdots \leq r_{(K)\mathbf{\Gamma}}$ be the ordered values of $r_1(\Gamma_1), \ldots, r_K(\Gamma_K)$. As I increases to infinity $c(I)^{-1} \log \overline{P}_{(l)\mathbf{a},\mathbf{\Gamma}} \to -r_{(K-l+1)\mathbf{\Gamma}}$ for $1 \leq l \leq K$ almost surely.

Fix k. From the above we note that $c(I)^{-1} \sum_{l=k}^{K} \log \overline{P}_{(l),\mathbf{a}_{l-1},\mathbf{\Gamma}} \to -\sum_{l=1}^{K-k+1} r_{(l)\mathbf{\Gamma}}$ almost surely. Choose $a < -\sum_{l=1}^{K-k+1} r_{(l)\mathbf{\Gamma}} < b$. We allow $a = -\infty$ and $-\infty < -\infty$. Consequently, for any $\epsilon > 0$ there exists I_ϵ such that for $I \geq I_\epsilon$, as $c(I) \to \infty$ when I increases to infinity, with probability at least $1 - \epsilon$ we get $a < c(I)^{-1} \sum_{j=k}^{K} \log \overline{P}_{(j)\mathbf{a},\mathbf{\Gamma}} < b$. For $I \geq I_\epsilon$ with probability at least

$1-\epsilon$

$$\Pr(\chi^2_{2(K-k+1)} > -2c(I)a) \le \Pr(\chi^2_{2(K-k+1)} > -2\sum_{j=k}^{K} \log \overline{P}_{(j)\mathbf{a},\mathbf{\Gamma}}) \le \Pr(\chi^2_{2(K-k+1)} > -2c(I)b).$$

Noting that, $\lim_{n\to\infty} n^{-1} \log \Pr(\chi^2_d > nx) = -x/2$ for any $x \ge 0$ and $d > 0$ we get

$$a \le \liminf_{I\to\infty} c(I)^{-1} \log \Pr(\chi^2_{2(K-k+1)} > -2\sum_{l=k}^{K} \log \overline{P}_{(l)\mathbf{a},\mathbf{\Gamma}}) \le \limsup_{I\to\infty} c(I)^{-1} \log \Pr(\chi^2_{2(K-k+1)} > -2\sum_{l=k}^{K} \log \overline{P}_{(l)\mathbf{a},\mathbf{\Gamma}}) \le b.$$

This is true for arbitrary $\epsilon > 0$ and arbitrary numbers a and b such that $a < -\sum_{l=1}^{K-k+1} r_{(l)\mathbf{\Gamma}} < b$. Thus we conclude that

$$\lim_{I\to\infty} c(I)^{-1} \log P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{ef}) = \lim_{I\to\infty} c(I)^{-1} \log \Pr(\chi^2_{2(K-k+1)} > -2\sum_{l=k}^{K} \log \overline{P}_{(l)\mathbf{a},\mathbf{\Gamma}}) = -\sum_{l=1}^{K-k+1} r_{(l)\mathbf{\Gamma}}.$$

This limit might be negative infinity.

Now consider $\log P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{g}) = \log g_k(\overline{P}_{(k)\mathbf{a},\mathbf{\Gamma}}, \ldots, \overline{P}_{(K)\mathbf{a},\mathbf{\Gamma}})$ for any $\mathfrak{g}$. From the assumption of the theorem we have $\Pr(g_k(U_k, \ldots, U_K) \le \alpha) \le \alpha$, for any $\alpha \in [0,1]$. Thus for any $0 \le x_k \le \cdots \le x_K$

$$g(x_k, \ldots, x_K) \ge \Pr(g_k(U_1, \ldots, U_{K-k+1}) \le g_k(x_k, \ldots, x_K)).$$

By the nondecreasing property of the function g_k

$$\Pr(g_k(U_1, \ldots, U_{K-k+1}) \le g_k(x_k, \ldots, x_K)) \ge \Pr(U_1 \le x_k, \ldots, U_K \le x_K) = \prod_{l=1}^{K-k+1} \Pr(U_l \le x_{j+k-1}) = \prod_{l=k}^{K} x_l.$$

Thus, $g(x_k, \ldots, x_K) \ge \prod_{l=k}^{K} x_l$. This implies

$$\log P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{g}) \ge \sum_{l=k}^{K} \log \overline{P}_{(l)\mathbf{a},\mathbf{\Gamma}}.$$

We get by dividing by $c(I)$ and taking the limit, for $1 \le k' \le k$

$$\liminf_{I\to\infty} c(I)^{-1} \log P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{g}) \ge \lim_{I\to\infty} c(I)^{-1} \sum_{l=k}^{K} \log \overline{P}_{(l)\mathbf{a},\mathbf{\Gamma}}$$
$$= -\sum_{l=1}^{K-k+1} r_{(l)\mathbf{\Gamma}} \ge -\sum_{l=1}^{K-k'+1} r_{(l)\mathbf{\Gamma}} = \lim_{I\to\infty} c(I)^{-1} \log P^{k'|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{e}\mathfrak{f}).$$

Proof of Proposition 7. Following the proof of Theorem 4 we have, for any $k = 1, \ldots, K$, $\lim_{I\to\infty} c(I)^{-1} \log P^{k|K}_{\mathbf{a},\mathbf{\Gamma}} = -\sum_{l=1}^{K-k+1} r_{(l)\mathbf{\Gamma}}$. Consider k such that $\Gamma_k > \widetilde{\Gamma}_k$. Since, $\widetilde{\Gamma}_k$ is the design sensitivity of the kth factor, by definition of the design sensitivity, $\overline{P}_{k,\mathbf{a}_{k-1},\Gamma_k} \to 1$. Further, since $c(I) \to \infty$ as $I \to \infty$, it implies $r_k(\Gamma_k) = 0$. The number of l with $\Gamma_l < \widetilde{\Gamma}_l$ is called $\widetilde{k}$. Hence, in the ordered values $r_{(1)\mathbf{\Gamma}} \le \cdots \le r_{(K)\mathbf{\Gamma}}$ the first $K - \widetilde{k}$ are zero. Thus the proof of the first part follows.

To prove of the final statement, consider the truncated product method. Let $\varkappa$ be the truncation level. For a number a let $a^{\varkappa}$ be the truncated version defined as a if $a < \varkappa$, otherwise it is 1. The combining method is $g_k(x_k, \ldots, x_K) = \Pr(\prod_{l=k}^{K} U_l^{\varkappa} < \prod_{l=k}^{K} x_l^{\varkappa})$, where $U_1, \ldots, U_K$ are i.i.d. uniform(0,1) random variables. For $0 \le x_1, \ldots, x_K \le 1$, let $y = -c(I)^{-1} 2 \log \prod_{l=k}^{K} x_l^{\varkappa}$. We write with $I \to \infty$ in mind (and K fixed)

$$\begin{aligned}
&g_k(x_k, \ldots, x_K) \\
&= \Pr(\prod_{l=k}^{K} U_l^{\varkappa} < \exp(-c(I)y/2)) \\
&= \Pr(-2\sum_{l=k}^{K} \log U_l^{\varkappa} > c(I)y) \\
&= \sum_{\mathcal{K}\subseteq\{k,\ldots,K\}} \Pr(-2\sum_{l=k}^{K} \log U_l^{\varkappa} > c(I)y \mid U_j \ge \varkappa, \forall j \in \mathcal{K}^c) \Pr(U_j \ge \varkappa, \forall j \in \mathcal{K}^c) \\
&= \sum_{\mathcal{K}\subseteq\{k,\ldots,K\}} \Pr(-2\sum_{l\in\mathcal{K}} \log U_l > c(I)y \mid U_j \ge \varkappa, \forall j \in \mathcal{K}^c) \Pr(U_j \ge \varkappa, \forall j \in \mathcal{K}^c) \\
&= \sum_{\mathcal{K}\subseteq\{k,\ldots,K\}} \Pr(-2\sum_{l\in\mathcal{K}} \log U_l > c(I)y) \Pr(U_j \ge \varkappa, \forall j \in \mathcal{K}^c) \\
&= \sum_{\mathcal{K}\subseteq\{k,\ldots,K\}, \mathcal{K}\neq\emptyset} \Pr(\chi^2_{2|\mathcal{K}|} > c(I)y) \times (1-\varkappa)^{|\mathcal{K}^c|}
\end{aligned}$$

$$= \sum_{\mathcal{K}\subseteq\{k,\ldots,K\},\mathcal{K}\neq\emptyset} \exp\{-c(I)y/2 + o(c(I))\} \times (1-\varkappa)^{|\mathcal{K}^c|}$$
$$= \exp\{-c(I)y/2 + o(c(I))\} \sum_{\mathcal{K}\subseteq\{k,\ldots,K\},\mathcal{K}\neq\emptyset} (1-\varkappa)^{|\mathcal{K}^c|}$$
$$= \exp\{-c(I)y/2 + o(c(I))\} \times \{1-(1-\varkappa)^{K-k+1}\}.$$

We used the fact that $\lim_{n\to\infty} n^{-1} \log \Pr(\chi^2_d > nx) = -x/2$ for any $x \geq 0$ and $d > 0$. Using the truncated product method (call it $\mathfrak{tp}$)

$$P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{tp}) = \exp\{-\log \prod_{l=k}^{K} \overline{P}^{\varkappa}_{(l)\mathbf{a},\mathbf{\Gamma}} + o(c(I))\} \times \{1-(1-\varkappa)^{K-k+1}\}.$$

Thus, $c(I)^{-1} \log P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{tp}) = \{-\sum_{l=k}^{K} c(I)^{-1} \log \overline{P}^{\varkappa}_{(l)\mathbf{a},\mathbf{\Gamma}} + o(1)\} + o(1)$. Finally, for large I, $\overline{P}^{\varkappa}_{(l)\mathbf{a},\mathbf{\Gamma}} = \overline{P}_{(l)\mathbf{a},\mathbf{\Gamma}}$ for all l since $\overline{P}_{(l)\mathbf{a},\mathbf{\Gamma}}$ converges to 0 or 1, in this setting. We get, from our proof of Theorem 4, $c(I)^{-1} \log P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{tp}) - c(I)^{-1} \log P^{k|K}_{\mathbf{a},\mathbf{\Gamma}}(\mathfrak{ef}) = o(1)$. This completes the proof.

More simulation results.

The following discussion supplements the simulation results of Section 6.3. The simulation settings are different from the ones reported in §6.3 in many ways. (1) We consider different sample sizes, $I = 200$ and $I = 500$. (2) $K = 4$. (3) We allow correlated treatments: for each unit, we simulated latent variables $x_1, \ldots, x_4$ from a multivariate normal with zero mean vector, with variance of the variables 1 and correlation of any two of them is 0.2; from that we defined $z_k = 1$ when $x_k < 0.1$. (4) The outcomes are simulated from $\chi^2_2/2 + \sum_k z_k\theta_k$. (5) We consider two treatment effect scenarios: Scenario 1: $\theta_k = 0$ for all k and Scenario 2: $\beta_k \sim Unif[0.1, 0.2]$.

The results of the simulation are reported in Table 4 for sample size $I = 200$ and in Table 5 for sample size $I = 500$. The comparative simulation results between the methods are similar to ones reported in §6.3. We make a few more observations based on these simulation results. In scenario 1, as the theory suggests, the family wise error rate is controlled at level 0.05 (5%). Increasing the sample size increases the power of the tests. Although, increasing the sample size does not increase the level of sensitivity to unmeasured confounding. As the sample size increases to infinity, there is a threshold of Γ, called the design sensitivity of the test, below that threshold the power goes to 1 and above it the power goes to 0.

TABLE 4

Simulation results for the power of sensitivity analysis evaluated at level 0.05. Numbers are out of 100. A cell value is the percentage of times the decision that at least k many $H_{0,l}$s are false is made, with $\Gamma_1 = \cdots = \Gamma_4 =: \Gamma$, out of 10000 simulations. Empty cells represent the value 0. tP = truncated product method with truncation level $\varkappa = 0.20$; sP = the modified additive p-value method in Proposition 4; Si = Simes' method; HB = Holm-Bonferroni method. **I= 200, K=4**

$k \rightarrow$	4				3				2				1			
$\Gamma \downarrow$	tP	sP	Si	HB	tP	sP	Si	HB	tP	sP	Si	HB	tP	sP	Si	HB
Scenario 1: (null case) $\beta_1 = \cdots = \beta_4 = 0$																
1	0	0	0	0	0	0	0	0	0	1	0	0	5	5	5	5
Scenario 2: $\beta_k \sim Unif(0.1, 0.2)$ for all k																
1	4	4	4	3	29	34	20	18	74	68	62	60	99	92	97	97
1.2	1	1	1	0	10	15	6	5	43	43	31	29	89	74	82	80
1.4					3	5	1	1	18	20	11	10	64	49	53	52
1.6					1	1			5	8	3	3	34	25	28	27
1.8									1	2	1	1	14	10	13	13
2													5	3	6	6
2.5															1	1
3																

TABLE 5

Simulation results for the power of sensitivity analysis evaluated at level 0.05. Numbers are out of 100. A cell value is the percentage of times the decision that at least k many $H_{0,l}$s are false is made, with $\Gamma_1 = \cdots = \Gamma_4 =: \Gamma$, out of 10000 simulations. Empty cells represent the value 0. tP = truncated product method with truncation level $\varkappa = 0.20$; sP = the modified additive p-value method in Proposition 4; Si = Simes' method; HB = Holm-Bonferroni method. **I= 500, K=4**

$k \rightarrow$	4				3				2				1			
$\Gamma \downarrow$	tP	sP	Si	HB	tP	sP	Si	HB	tP	sP	Si	HB	tP	sP	Si	HB
Scenario 1: (null case) $\beta_1 = \cdots = \beta_4 = 0$																
1	0	0	0	0	0	0	0	0	0	1	0	0	5	5	5	5
Scenario 2: $\beta_k \sim Unif(0.1, 0.2)$ for all k																
1	24	24	24	22	72	70	64	61	98	92	95	94	100	99	100	100
1.2	7	7	7	5	36	39	27	26	83	73	74	72	100	94	99	99
1.4	1	1	1	1	12	15	7	7	48	42	39	37	93	74	89	88
1.6					2	4	1	1	16	16	12	12	64	42	60	58
1.8									3	4	2	2	28	14	28	28
2													8	3	10	10
2.5																
3																

References

Bahadur, R. R. (1967). Rates of convergence of estimates and test statistics. *Annals of Mathematical Statistics*, **38** 303–324.

Becker, B. J. (1994). Combining significance levels. In Cooper, H and Hedges, L V, editors *A handbook of research synthesis*, chapter 15, pages 215–230, Russell Sage, New York, 1994.

Benjamini, Y. and Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics*, **64(4)** 1215–1222.

Benjamini, Y., Heller, R. and Yekutieli, D. (2009). Selective inference in complex research. *Philosophical Transaction of the Royal Society A*, **367**, 4255–4271.

Centerwall, B. S. (1989). Exposure to television as a risk factor for violence. *American Journal of Epidemiology*, **129(4)**, 643–652.

Chen, Z. (2011). Is the weighted z-test the best method for combining probabilities from independent tests? *Journal of Evolutionary Biology* **24** 926–930.

Cochran, W. G. (1965). The planning of observational studies in human population (with Discussion). *Journal of the Royal Statistical Society A*, 234–266.

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

Crupi, V., Charter, N. and Tentori, K. (2013). New axioms for probability and likelihood ratio measures. *The British Journal of Philosophy of Science*, **64** 189–204.

Ding, P. and Vanderweele, T. J. (2016). Sensitivity analysis without assumptions. *Epidemiology*, **27** 368–377.

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *The Journal of Psychology*, **80** 351–363.

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

Ertefaie, A., Small, D. S. and Rosenbaum, P. R. (2018). Quantitative evaluation of the trade-off of strengthened instruments and sample size in observational studies. *Journal of the American Statistical Association*, **113(523)** 1122–1134.

Fisher, R. A. (1932). *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

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

Fogarty, C. B. and Hasegawa, R. B. (2018). Extended sensitivity analysis for heterogeneous unmeasured confounding with an application to sibling studies of returns to education. *Annals of Applied Statistics*, *to appear*.

Fogarty, C. B. and Small, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *Journal of the American Statistical Association*, **111** 1820–1830.

Gilbert, P., Bosch, R., Hudgens, M. (2003). Sensitivity analysis for the assessment of the causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*, **59** 531–541.

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

Hansen, B. B., Rosenbaum, P. R. and Small, D. S. (2014). Clustering treatment assignments and sensitivity to unmeasured biases in observational studies. *Journal of the American Statistical Association*, **109** 133–144.

Helland, I. S. (1995). Simple counterexamples against the conditionality principle. *The American Statistician* **49** 351–356.

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

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

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

Kalbfleisch, J. D. (1975). Sufficiency and conditionality. *Biometrika*, **62** 251–259.

Keele, L. and Minozzi, W. (2013). How much is Minnesota like Wisconsin? Assumptions and conterfactuals in causal inference with observational data. *Political Analysis*, **21** 193–216.

Lancaster, H. (1961). The combination of probabilities: an application of orthonormal functions. *Australian & New Zealand Journal of Statistics*, **3** 20–33.

Lipták, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Közl.*, **3** 171–196.

Liu, W., Kuramoto, J. and Stuart, E. (2013). Sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science: the official journal of the Society for Prevention Research*, **14** 570–580.

Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M. and Saah, M. (1982). Lead absorption in children of emplyees in a lead-related factory. *American Journal of Epidemiology*, **115** 549–555.

Neyman, J. (1923, 1990). On the application of probability theory to agricultural experiments. *Statistical Science*, **5** 463–480.

Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association*, **110** 515–527.

Popper, K. R., Sir (1934). *Logik der Forschung*, Julius Springer Verlag, Vienna.

Popper, K. R., Sir (1954). Degree of confirmation. *The British Journal of Philosophy of Science*, **5** 143–149.

Popper, K. R., Sir (1972). *The Logic of Scientific Discovery (6th impression revised)*, Hutchinson, London.

Popper, K. R., Sir (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*, New York: Routledge and Kegan Paul.

Reynolds, K. D. and West, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs, *Evaluation Review*, **11** 691–714.

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

Rosenbaum, P. R. (2002). *Observational Studies* (2nd edition), New York: Springer.

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

Rosenbaum, P. R. (2005). Observational Study. In *Encyclopedia of Statistics in Behavioral Science*, Editors Brian S. Everitt & David C. Howell, Volume 3, pp. 1451–1462

Rosenbaum, P. R. (2010). Design sensitivity and efficiency in observational studies. *Journal of the American Statistical Association*, **105(490)** 692–702.

Rosenbaum, P. R. (2011). Some approximate evidence factors in observational studies. *Journal of the American Statistical Association*, **106(493)** 285–295.

Rosenbaum, P. R. (2015). Bahadur efficiency of sensitivity analyses in observational studies. *Journal of the American Statistical Association*, **110** 205–217.

Rosenbaum, P. R. (2017). The general structure of evidence factors in observational studies. *Statistical Science*, **32(4)** 514–530.

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

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

Rowbottom, D. P. (2013). Popper's measure of corroboration and P(h|b). *The British Journal of Philosophy of Science*, **64** 739–745.

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

Shaked, M. and Shanthikumar, G. (2007). *Stochastic Orders*, Springer Series in Statistics, Springer-Verlag New York.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73** 751–754.

Sprenger, J. (2018). Two impossibility results for measure of corroboration. *The British Journal of Philosophy of Science*, **69** 139–159.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A. and Williams, R. M. Jr. (1949). *The American Soldier, Vol.1: Adjustment during Army Life*, Princeton University Press, Princeton,.

Stuart, E. A., DuGoff, E., Abrams, M., Salkever, D. and Steinwachs, D. (2013). Estimating causal effects in observational studies using electronic health data: challenges and (some) solutions. *eGEMs*, **1** Article 4.

Wong, M., Cook, T. D. and Steiner, P. M. (2015). Adding design elements to improve time series designs: No child left behind as an example of causal pattern-matching. *Journal of Research on Educational Effectiveness*, **8** 245–279.

Wang, J. and Owen, A. B. (2017). Admissibility in partial conjunction testing. *Journal of the American Statistical Association*, to appear.

Whitlock, M. C. (2005). Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, **18** 1368–1373.

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

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., & Weir, B. S. (2002). Truncated product method for combining P-values. *Genetic Epidemiology*, **22(2)** 170–185.

Zaykin, D. V. (2011). Optimally weighted Z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology* **24(8)** 1836–1841.

Zhao, Q. (2018). On sensitivity value of pair-matched observational studies. *Journal of the American Statistical Association*, to appear.

Zubizarreta, J. R., Neuman, M., Silber, J. H. & Rosenbaum, P. R. (2012). Contrasting evidence within and between institutions that provide treatment in an observational study of alternate forms of anesthesia. *Journal of the American Statistical Association*, **107(499)** 901-915.

Zubizarreta, J. R., Paredes, R. D. & Rosenbaum, P. R. (2014). Matching for balance, pairing for heterogeneity in an observational study of the effectiveness of for-profit and not-for-profit high schools in Chile. *Annals of Applied Statistics*, **8** 204–231.

Department of Statistics,
The Wharton School,
University of Pennsylvania
400 Jon M. Huntsman Hall
3730 Walnut Street
Philadelphia, PA 19104-6340
E-mail: bikramk@wharton.upenn.edu
dsmall@wharton.upenn.edu