The Annals of Applied Statistics
2016, Vol. 10, No. 3, 1447–1471
DOI: 10.1214/16-AOAS942

USING SCHEFFÉ PROJECTIONS FOR MULTIPLE OUTCOMES IN AN OBSERVATIONAL STUDY OF SMOKING AND PERIODONTAL DISEASE

BY PAUL R. ROSENBAUM

University of Pennsylvania

In an observational study of the effects caused by treatments, a sensitivity analysis asks about the magnitude of bias from unmeasured covariates that would need to be present to alter the conclusions of a naive analysis that presumes adjustments for measured covariates remove all biases. When there are two or more outcomes in an observational study, these outcomes may be unequally sensitive to unmeasured biases, and the least sensitive finding may concern a combination of several outcomes. A method of sensitivity analysis is proposed using Scheffé projections that permits the investigator to consider all linear contrasts in two or more scored outcomes while controlling the family-wise error rate. In sufficiently large samples, the method will exhibit insensitivity to bias that is greater than or equal to methods, such as the Bonferroni–Holm procedure, that focus on individual outcomes; that is, Scheffé projections have larger design sensitivities. More precisely, if the least sensitive linear combination is a single one of the several outcomes, then the design sensitivity using Scheffé projections equals that using a Bonferroni correction, but if the least sensitive combination is a nontrivial combination of two or more outcomes, then Scheffé projections have larger design sensitivities. This asymptotic property is examined in terms of finite sample power of sensitivity analyses using simulation. The method is applied to a replication with recent data of a well-known study of the effects of smoking on periodontal disease. In the example, the comparison that is least sensitive to bias from unmeasured covariates combines results for lower and upper teeth, but emphasizes lower teeth. This pattern would be difficult to anticipate prior to examining the data, but Scheffé's method permits use of this unanticipated pattern without fear of capitalizing on chance.

1. Introduction: Motivating example; outline.

1.1. *Smoking and periodontal disease.* Cigarette smoking is widely believed to be a cause of periodontal disease. Using data from NHANES III, Tomar and Asma (2000) claimed that about 42% of cases of periodontal disease in the US are attributable to smoking. A comparison of this kind entails consideration of potential biases. In the US, smoking is more common among people with less education and less income, who may have reduced access to professional dental

Received August 2015; revised January 2016.
Key words and phrases. Causal inference, design sensitivity, observational study, Scheffe projection, sensitivity analysis.

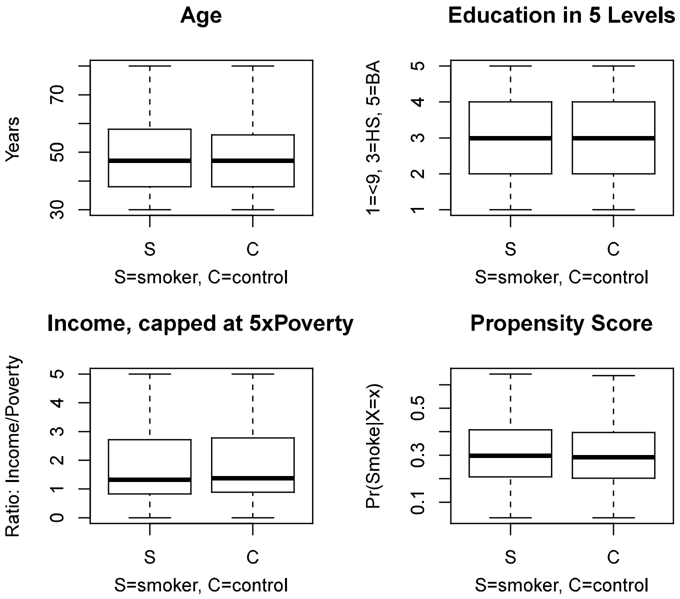

FIG. 1. *Three covariates and a propensity score in $I = 441$ matched pairs of daily smokers* (*S*) *and never smokers* (*C*) *from NHANES* 2011–2012.

care. Periodontal disease increases dramatically with age, and fewer smokers live to advanced ages. Smoking expresses a lack of concern with health that might also be manifested in many other ways, perhaps including poor personal dental care.

Using more recent data from NHANES 2011–2012, Figure 1 and Table 1 show 441 matched pairs of a daily smoker and a never smoker. Daily smokers smoked

TABLE 1

Covariates in matched and unmatched samples. Smokers have less education and income, are younger, less often female, more often black. Education is 1–5 *with* 1 = 9*th grade*, 3 = *High School*, 5 = *College Degree*

Covariate	Treated smoker	Controls: Never smokers		
		Matched	Unmatched	All
Sample size	441	441	1065	1506
No High School Degree %	29	29	14	18
Education (mean)	3.2	3.2	4.0	3.7
Income/(Poverty Level)	1.9	2.0	3.2	2.9
Age, mean	48	48	53	52
Age ≥60%	18	18	35	30
Female %	40	42	67	60
Black %	32	31	22	25

every day of the last 30 days. Never smokers smoked fewer than 100 cigarettes in their life, do not smoke now, and had no tobacco use in the previous five days. Attention is restricted to the subsample of NHANES who were given and completed a periodontal exam and had at least one periodontal measurement. Pairs are matched for education, income (recorded as a ratio, a multiple of the poverty level, capped at five times poverty), age, gender and black race. In Table 1, before matching, smokers have less education and income, are younger, more often male and more often black. In Table 1 and Figure 1, after matching, these visible differences have been removed, but of course the groups may differ in other ways not recorded by NHANES. For discussion of multivariate matching, see Hansen (2007), Stuart (2010) and Zubizarreta (2012).

Helpful diagrammatic descriptions of periodontal measurements in NHANES are given by Wei, Barker and Eke (2013). Measurements are made for 28 teeth, 14 upper and 14 lower teeth, excluding 4 wisdom teeth. Pocket depth and loss of attachment are two complementary measures of the degree to which the gums have separated from the teeth. Pocket depth and loss of attachment are measured in six locations on each tooth, providing the tooth is present. In parallel with Tomar and Asma (2000), a periodontal measurement at a location was taken to exhibit disease if it had either a loss of attachment of ≥ 4 mm or a pocket depth of ≥ 4 mm, so each tooth contributes a score of 0-to-6. By this definition, Figure 2 depicts for each person the proportion of measurements exhibiting periodontal disease, for upper and lower teeth. The smoker-minus-control differences are somewhat larger but also somewhat more unstable for lower teeth.

Figure 2 is the simplest example of an observational study with a bivariate outcome. For each outcome, there appears to be a substantial difference between smokers and nonsmokers, but how sensitive are these differences to unmeasured biases? Here are two possible analyses that one might select without giving the matter much thought. First, one might combine 14 lower teeth and 14 upper teeth into 28 teeth, declining to consider a bivariate outcome. In this case, one would do a sensitivity analysis for univariate matched pairs. Second, one might do two univariate sensitivity analyses, one for upper teeth, one for lower teeth, applying the Bonferroni–Holm procedure to the two bounds on P-values to correct for multiple testing, as in Rosenbaum and Silber (2009b), Section 4.5; however, see Fogarty and Small (2016) for a clever alternative approach improving this Bonferroni–Holm technique. For the long-tailed data in Figure 2, analyses might reasonably use a robust procedure, such as a rank test or an M-test, both of which score the observations in such a way as to limit the influence of extreme outliers.

An alternative procedure proposed in the current paper uses Scheffé projections; that is, it considers all possible linear combinations of scored outcomes, correcting using Scheffé's (1953) argument for multiple testing. Scheffé's argument is most familiar in the context of one-way analysis of variance, but it is applicable to multivariate outcomes. Section 3 develops a new sensitivity analysis for unmeasured biases with multivariate outcomes whose justification combines a minimax

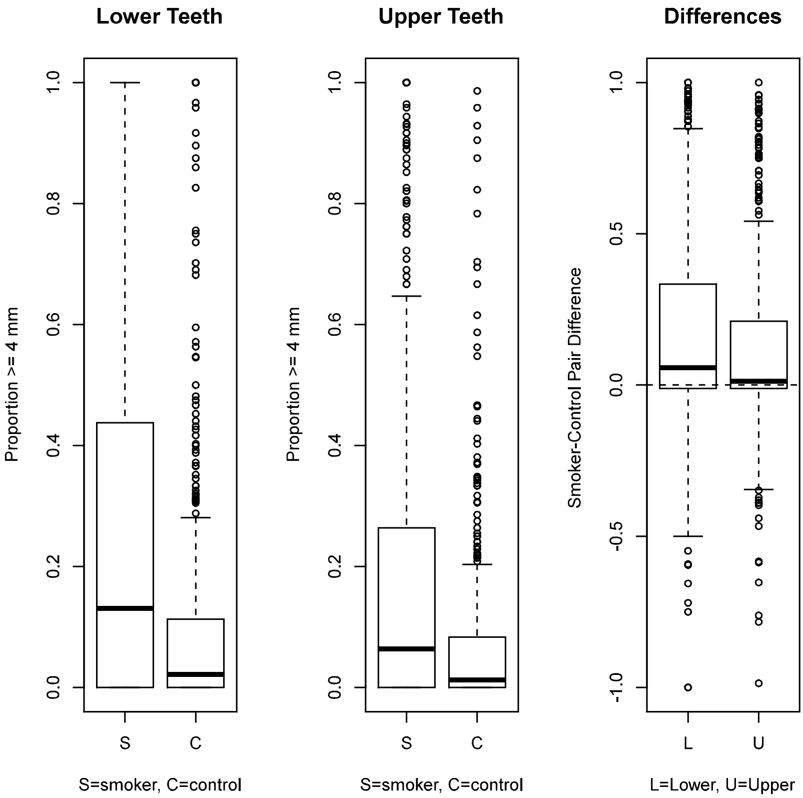

FIG. 2. *For* 441 *daily smokers and* 441 *matched never smokers*, *for lower teeth and upper teeth*, *boxplots depict the proportion of measurements exhibiting either a pocket depth of* 4 *mm or more or a loss of attachment of* 4 *mm or more*. *The plots show either* 441 *smokers and* 441 *controls*, *or* 441 *smoker-minus-control matched pair differences*.

inequality with Scheffé's argument. How does the sensitivity of the best linear combination compare to the analyses in the previous paragraph?

1.2. *Outline*: *A new procedure*, *its power and design sensitivity*. Section 2 is a brief review of causal inference and sensitivity analysis. Notation is reviewed in Section 2.1, randomization inference in Section 2.2 and sensitivity analyses in Section 2.3. Proposition 1 in Section 3.2 is one of the main results: it proposes a method of sensitivity analysis using any linear contrast that the investigator selects having examined the data. The proof of Proposition 1 combines a minimax inequality and the idea underlying Scheffé projections. Section 4 evaluates the performance of the proposed method in terms of the power of a sensitivity analysis and its asymptotic index, the design sensitivity. In Section 4.2, numerical calculations are given for design sensitivity in the case of bivariate outcomes in matched pairs; here, combinations of two outcomes can be more insensitive to unmeasured biases then either component. Then Section 4.3 shows that picking the best contrast empirically yields the same design sensitivity as knowing a priori the optimal contrast. Results in Sections 4.1–4.3 are asymptotic. In Section 4.4, the finite sample power is determined by simulation, with some loss of power due to correcting for repeated use of the data, but also some large gains in power if the best contrast is a nontrivial combination of two outcomes.

2. Notation, background, review.

2.1. *Notation*: *Treatment effects and treatment assignments.* There are I matched sets, $i = 1, \dots, I$, and set i contains $n_i \geq 2$ subjects, $j = 1, \dots, n_i$, so that ij refers to a particular person, the jth person in set i. Each matched set i contains one treated subject with $Z_{ij} = 1$, and $n_i - 1$ untreated controls with $Z_{ij} = 0$, so $1 = \sum_{j=1}^{n_i} Z_{ij}$ for each i. There are $N = \sum_{i=1}^{I} n_i$ subjects in total. The matching controlled an observed covariate $\mathbf{x}_{ij}$, so $\mathbf{x}_{i1} = \cdots = \mathbf{x}_{i,n_i}$ for each i, but subjects in the same matched set may differ in terms of an unobserved covariate u_{ij}. Each subject has two potential K-dimensional vector responses: the response, $\mathbf{r}_{Tij}$, that would be observed if subject ij were assigned to treatment with $Z_{ij} = 1$, and the response, $\mathbf{r}_{Cij}$, that would be observed from this same subject if assigned to control with $Z_{ij} = 0$; therefore, the response observed from subject ij is $\mathbf{R}_{ij} = Z_{ij}\mathbf{r}_{Tij} + (1 - Z_{ij})\mathbf{r}_{Cij}$, and the effect caused by the treatment, $\mathbf{r}_{Tij} - \mathbf{r}_{Cij}$, is not observed for any subject; see Neyman (1923, 1990), Welch (1937) and Rubin (1974). In later sections, $\mathbf{r}_{Tij}$ and $\mathbf{r}_{Cij}$ are each bivariate responses describing the lower and upper teeth for subject ij if this subject smokes daily, $\mathbf{r}_{Tij}$, or is a lifelong nonsmoker, $\mathbf{r}_{Cij}$, so the effect of daily smoking is $\mathbf{r}_{Tij} - \mathbf{r}_{Cij}$. Fisher's (1935) sharp null hypothesis of no treatment effect asserts $H_0 : \mathbf{r}_{Tij} = \mathbf{r}_{Cij}$ for all ij, that is, it asserts that different people ij have different degrees of periodontal health $\mathbf{r}_{Cij}$; however, $\mathbf{r}_{Cij}$ is not altered by smoking. Write $\mathcal{F} = \{(\mathbf{r}_{Tij}, \mathbf{r}_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, \dots, I, j = 1, \dots, n_i\}$. In Fisher's theory of randomization inference in randomized experiments, only the treatment assignment Z_{ij} and quantities like $\mathbf{R}_{ij}$ that depend on Z_{ij} are random variables, and $\mathcal{F}$ is fixed by conditioning.

Write $\mathbf{Z} = (Z_{11}, Z_{12}, \dots, Z_{I,n_I})^T$ and $\mathbf{u} = (u_{11}, u_{12}, \dots, u_{I,n_I})^T$ for the N-dimensional vectors, and $\mathbf{R}$, $\mathbf{r}_T$, $\mathbf{r}_C$ for the corresponding $N \times K$ matrices containing the $\mathbf{R}_{ij}$, $\mathbf{r}_{Cij}$, $\mathbf{r}_{Tij}$ as rows in the lexical order. For a finite set $\mathcal{S}$, write $|\mathcal{S}|$ for the number of elements of $\mathcal{S}$. Write $\mathcal{Z}$ for the set containing the $|\mathcal{Z}| = \prod_{i=1}^{I} n_i$ possible values $\mathbf{z}$ of $\mathbf{Z}$, so $\mathbf{z} \in \mathcal{Z}$ if $\mathbf{z} = (z_{11}, \dots, z_{I,n_I})^T$ with $z_{ij} = 0$ or $z_{ij} = 1$ and $1 = \sum_{j=1}^{n_i} z_{ij}$ for each i. Conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ is abbreviated as conditioning on $\mathcal{Z}$.

2.2. *Randomization inference in randomized experiments.* In a matched randomized experiment, $\mathbf{Z}$ is picked at random with equal probabilities from $\mathcal{Z}$ so that

$\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = |\mathcal{Z}|^{-1}$ for each $\mathbf{z} \in \mathcal{Z}$. A key part of randomization is that treatment assignment probabilities conditionally given $\mathcal{F}$—namely, $\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z})$—do not depend on $\mathcal{F}$; that is, the coin flips are truly fair in ignoring aspects of people recorded in $\mathcal{F}$. A test statistic T is a function of observed responses, $\mathbf{R}$, and observed treatment assignments, $\mathbf{Z}$; that is, $T = t(\mathbf{Z}, \mathbf{R})$. For instance, with matched pairs, the mean treated-minus-control pair difference is a test statistic, $T = t(\mathbf{Z}, \mathbf{R})$; indeed, it is the simplest of the M-statistics used in the current paper. If Fisher's hypothesis H_0 of no treatment effect is true, then $\mathbf{R} = \mathbf{r}_C$, where $\mathbf{r}_C$ is in $\mathcal{F}$ and hence is fixed conditionally given $\mathcal{F}$. In a randomized experiment, the distribution of treatment assignments $\mathbf{Z}$ is uniform on $\mathcal{Z}$; that is, $\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = |\mathcal{Z}|^{-1}$ for each $\mathbf{z} \in \mathcal{Z}$. Putting this together, if Fisher's null hypothesis H_0 of no effect is true in a randomized experiment, then the distribution of $T = t(\mathbf{Z}, \mathbf{R})$ is its permutation distribution,

$$\begin{aligned}\Pr\{t(\mathbf{Z}, \mathbf{R}) \geq v|\mathcal{F}, \mathcal{Z}\} &= \Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq v|\mathcal{F}, \mathcal{Z}\} \\ &= \frac{|\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq v\}|}{|\mathcal{Z}|},\end{aligned} \tag{2.1}$$

because $\mathbf{R} = \mathbf{r}_C$ if H_0 is true, $\mathbf{r}_C$ is fixed by conditioning on $\mathcal{F}$, and $\mathbf{Z}$ is uniform on $\mathcal{Z}$ in a randomized experiment. In (2.1), the tail probability $\Pr\{t(\mathbf{Z}, \mathbf{R}) \geq v|\mathcal{F}, \mathcal{Z}\}$ is simply the proportion of treatment assignments $\mathbf{z} \in \mathcal{Z}$ that would produce a value of $t(\mathbf{z}, \mathbf{R})$ of v or more under H_0 with $\mathbf{R}$ not changing as $\mathbf{z}$ changes. If $T = t(\mathbf{Z}, \mathbf{R})$ is the mean of I treated-minus-control pair differences, then the distribution (2.1) of T under H_0 in a randomized experiment—the so-called permutational t-test—is found by computing the mean for all 2^I possible changes in signs of the I pair differences; see Lehmann and Romano (2005), Section 5, for general discussion, see Maritz (1979) for the case of M-statistics, and see Rosenbaum (2010), Section 2.9, for a tiny example presented in explicit detail.

2.3. *Model for sensitivity analysis in observational studies.* The randomization distribution in (2.1) is derived from the random assignment of treatments in an experiment, and there is typically no reason to believe it is applicable in an observational or nonrandomized study of treatment effects. A simple model for treatment assignment in observational studies says that in the population before matching, treatments are assigned independently with unknown probabilities $\pi_{ij} = \Pr(Z_{ij} = 1|\mathcal{F})$ such that two subjects, ij and ij', who might be matched because they have the same value of the observed covariates, $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$, may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$, that is,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1 - \pi_{ij'})}{\pi_{ij'}(1 - \pi_{ij})} \leq \Gamma \qquad \text{whenever } \mathbf{x}_{ij} = \mathbf{x}_{ij'}, \tag{2.2}$$

and then returns the distribution of $\mathbf{Z}$ to $\mathcal{Z}$ by conditioning on $\mathbf{Z} \in \mathcal{Z}$. Writing $\mathcal{U} = [0, 1]^N$ for the N-dimensional unit cube and $\gamma = \log(\Gamma)$, it is easy to verify [Rosenbaum (1995)] that this model is equivalent to assuming that, for $\mathbf{z} \in \mathcal{Z}$,

$$\Pr(\mathbf{Z} = \mathbf{z}|\mathcal{F}, \mathcal{Z}) = \prod_{i=1}^{I} \frac{\exp(\gamma \sum_{j=1}^{n_i} z_{ij} u_{ij})}{\sum_{j=1}^{n_i} \exp(\gamma u_{ij})}, \qquad \text{with } \mathbf{u} \in \mathcal{U}, \tag{2.3}$$

where the verification consisting in constructing u_{ij} satisfying (2.3) from π_{ij} satisfying (2.2) and conversely.

Because π_{ij} and u_{ij} are unknown, the distributions in (2.2) and (2.3) are unknown, so for several values of $\Gamma \geq 1$ a sensitivity analysis computes bounds on inference quantities such as P-values or point estimates, thereby determining the magnitude of bias Γ in treatment assignment that would need to be present to alter the qualitative conclusions of an observational study. In principle, an exact computation of $\Pr\{t(\mathbf{Z}, \mathbf{R}) \geq v|\mathcal{F}, \mathcal{Z}\}$ under H_0 for fixed $\gamma = \log(\Gamma)$ and $\mathbf{u} \in \mathcal{U}$ entails summing terms (2.3) over $\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{r}_C) \geq v\}$, yielding (2.1) for $\gamma = 0$; then a sensitivity bound is obtained by maximizing and minimizing $\Pr\{t(\mathbf{Z}, \mathbf{R}) \geq v|\mathcal{F}, \mathcal{Z}\}$ over $\mathbf{u} \in \mathcal{U}$. This exact calculation is feasible for moderate I for matched pairs, $n_i = 2$ for some test statistics [Rosenbaum (2010), Section 3.9]. If, as is often true, the test statistic is of the form $T = \sum_{i=1}^{I} \sum_{j=1}^{n_i} Z_{ij} q_{ij}$ where q_{ij} is a function of $\mathbf{R}$ (and hence a function of $\mathbf{r}_C$ under H_0), then for fixed γ and $\mathbf{u}$, under H_0 and mild conditions on the scores q_{ij}, the distribution of $\{T - \mathrm{E}_{\Gamma,\mathbf{u}}(T)\}/\sqrt{\mathrm{var}_{\Gamma,\mathbf{u}}(T)}$ under (2.3) converges to the Normal distribution as $I \to \infty$, where $\mathrm{E}_{\Gamma,\mathbf{u}}(T)$ and $\mathrm{var}_{\Gamma,\mathbf{u}}(T)$ are the expectation and variance of T under (2.3). In this case, a simple algorithm calculates approximate bounds on $\Pr\{t(\mathbf{Z}, \mathbf{R}) \geq v|\mathcal{F}, \mathcal{Z}\}$ for fixed $v > 0$ using this Normal approximation for a particular $\mathbf{u} \in \mathcal{U}$ that maximizes $\mathrm{E}_{\Gamma,\mathbf{u}}(T)$ and maximizes $\mathrm{var}_{\Gamma,\mathbf{u}}(T)$ among all $\mathbf{u}$ that maximize $\mathrm{E}_{\Gamma,\mathbf{u}}(T)$; see Gastwirth, Krieger and Rosenbaum (2000) and Rosenbaum (2007, 2014), and see the R packages sensitivitymv (version 1.3) and sensitivitymw (version 1.1) for implementation in the case of M-statistics including the mean. This algorithm is called a "separable approximation" because $\mathrm{E}_{\Gamma,\mathbf{u}}(T)$ and $\mathrm{var}_{\Gamma,\mathbf{u}}(T)$ can be optimized separately, one matched set at a time, and then combined, whereas the exact tail probability cannot be optimized in this way. In the current paper, when a deviate $|T - \mathrm{E}_{\Gamma,\mathbf{u}}(T)|/\sqrt{\mathrm{var}_{\Gamma,\mathbf{u}}(T)}$ is minimized over $\mathbf{u} \in \mathcal{U}$, the calculation uses the separable approximation to that minimum. Simpler algorithms work with matched pairs, $n_i = 2$; see Rosenbaum (2007).

For discussion of various methods, aspects and illustrations of sensitivity analyses in observational studies, see Cornfield et al. (1959), Rosenbaum and Rubin (1983), Manski (1990), Manski and Nagin (1990), Yu and Gastwirth (2005), Shepherd et al. (2006), McCandless, Gustafson and Levy (2007), Heller, Rosenbaum and Small (2009), Hosman, Hansen and Holland (2010), Hsu and Small (2013) and Liu, Kuramoto and Stuart (2013). In particular, in some simple situations, bounds related to Manski's bounds are obtained by letting $\Gamma \to \infty$ in (2.2); see Rosenbaum (1995), Section 2.4.

3. Comparisons among multiple outcomes.

3.1. *Weighted combinations of test statistics for several outcomes.* The K outcomes are examined with L test statistics of the form $T_\ell = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Z_{ij} q_{ij\ell}$ where $q_{ij\ell}$ is a function of $\mathbf{R}$, $\ell = 1, \dots, L$. Commonly, $L = K$, and q_{ijk} will be a function of the entries in the kth column of $\mathbf{R}$, that is, of the kth outcome variable, but this is not essential and there is no gain in assuming this when testing Fisher's null hypothesis of no treatment effect $H_0 : \mathbf{r}_T = \mathbf{r}_C$. Under Fisher's H_0, $\mathbf{R} = \mathbf{r}_C$ is fixed by conditioning on $\mathcal{F}$ in (2.1) and (2.3), so $q_{ij\ell}$ is also fixed, and T_ℓ is the sum of the fixed scores $q_{ij\ell}$ for those individuals assigned to treatment, $Z_{ij} = 1$. Write $\mathbf{T} = (T_1, \dots, T_L)^T$ for the L-dimensional vector.

Many familiar statistics have the form $T_k = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Z_{ij} q_{ijk}$ when $K = L$ and T_k refers to the kth of the K outcomes R_{ijk}. Among rank tests, taking q_{ijk} to be the rank of R_{ijk} among the n_i individuals in set i makes T_k into a stratified Wilcoxon rank sum statistic, whereas taking $(n_i + 1)q_{ijk}$ to be the rank of R_{ijk} yields an optimally weighted combination of rank sum tests [van Elteren (1960); Lehmann (1975), Section 3.3, page 135]. If q_{ijk} is a rank of $R_{ijk} - n_i^{-1} \sum_{j'=1}^{n_i} R_{ij'k}$ ranking from 1 then N, then T_k is Hodges and Lehmann's (1962) aligned rank statistic for the kth outcome. Define $D_{ijj'k} = R_{ijk} - R_{ij'k}$, so $D_{ijjk} = 0$ and $D_{ijj'k} = r_{Cijk} - r_{Cij'k}$ if H_0 is true. If $q_{ijk} = \sum_{j'=1}^{n_i} D_{ijj'k}/\{I(n_i - 1)\}$, then $T_k = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Z_{ij} q_{ijk}$ is the mean over I matched sets of the mean treated-minus-control difference in the kth outcome within set i, and it is an unbiased estimate of the average effect of the treatment on treated subjects in a matched randomized experiment; moreover, for pairs with $n_i = 2$, this T_k is familiar as the basis for the permutational t-test of Fisher (1935), Pitman (1937) and Welch (1937). Another test statistic is based on Huber's M-statistics; see Maritz (1979) for randomization inference using M-statistics when $n_i = 2$. Let $\psi(\cdot)$ be an odd function, $\psi(y) = -\psi(-y)$, so that $\psi(0) = 0$. Huber (1981) favored $\psi_{\text{hu}}(y) = \text{sign}(y)\min(|y|, \kappa)$ for some $\kappa > 0$, while $\psi_{\text{t}}(y) = y$ again yields a permutational t-test for $n_i = 2$. There are $\sum_{i=1}^{I}\binom{n_i}{2}$ values of $|D_{ijj'k}|$ with $j < j$, so define s_k to be a quantile, typically the median of $|D_{ijj'k}|$. In the example in Section 1.1 and Section 5, and in the numerical results in Section 4, $\kappa = 2.5$ and s_k is the median absolute pair difference. Let $w_{ik} \geq 0$ be a weight that is a function of $\mathbf{R}$, I and the n_i, most commonly, $w_{ik} = 1$. Then $T_k = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Z_{ij} q_{ijk}$ is a weighted M-statistic if $q_{ijk} = w_{ik} \sum_{j'=1}^{n_i} \psi(D_{ijj'k}/s_k)$; see Rosenbaum (2007, 2013, 2014) for detailed discussion, including sensitivity analyses for M-tests, M-estimates and confidence intervals, and for the relative performance of different $\psi(\cdot)$ functions and weights w_i, and see the R packages `sensitivitymv` (version 1.3) and `sensitivitymw` (version 1.1) for implementation.

Under H_0 and (2.3) for fixed $\gamma = \log(\Gamma)$ and $\mathbf{u} \in \mathcal{U}$,

$$\mathrm{E}_{\Gamma,\mathbf{u}}(T_\ell | \mathcal{F}, \mathcal{Z}) = \sum_{i=1}^{I}\sum_{j=1}^{n_i} \frac{q_{ij\ell}\exp(\gamma u_{ij})}{\sum_{j=1}^{n_i}\exp(\gamma u_{ij})} = \mu_{\ell,\mathbf{u}}, \qquad \text{say,}$$

and

$$\mathrm{cov}_{\Gamma,\mathbf{u}}(T_\ell, T_{\ell'}|\mathcal{F},\mathcal{Z}) = \sum_{i=1}^{I}\sum_{j=1}^{n_i}\frac{q_{ij\ell}q_{ij\ell'}\exp(\gamma u_{ij})}{\sum_{j=1}^{n_i}\exp(\gamma u_{ij})} - \mu_{\ell,\mathbf{u}}\mu_{\ell',\mathbf{u}} = \sigma_{\ell,\ell',\mathbf{u}}, \qquad \text{say.}$$

Write $\boldsymbol{\mu}_{\mathbf{u}}^{(\Gamma)} = (\mu_{1,\mathbf{u}},\dots,\mu_{L,\mathbf{u}})^T$ and $\boldsymbol{\Sigma}_{\mathbf{u}}^{(\Gamma)}$ for the $L \times L$ matrix containing the $\sigma_{\ell,\ell',\mathbf{u}}$, where $\boldsymbol{\Sigma}_{\mathbf{u}}^{(\Gamma)}$ is assumed to be positive definite. When there is no chance of confusion, it is less cumbersome to write $\boldsymbol{\mu}_{\mathbf{u}}$ for $\boldsymbol{\mu}_{\mathbf{u}}^{(\Gamma)}$ and $\boldsymbol{\Sigma}_{\mathbf{u}}$ for $\boldsymbol{\Sigma}_{\mathbf{u}}^{(\Gamma)}$, leaving implicit the dependence of $\boldsymbol{\mu}_{\mathbf{u}}$ and $\boldsymbol{\Sigma}_{\mathbf{u}}$ on Γ. In a few places, the explicit if cumbersome notation is needed, hence used.

Finally, let $\boldsymbol{\lambda} = (\lambda_1,\dots,\lambda_L)^T$ be an L-dimensional vector of constants and define

$$(3.1)\qquad \begin{aligned} T_{\boldsymbol{\lambda}} &= \sum_{\ell=1}^{L}\lambda_\ell T_\ell = \boldsymbol{\lambda}^T\mathbf{T} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Z_{ij}\sum_{\ell=1}^{L}\lambda_\ell q_{ij\ell} \\ &= \sum_{i=1}^{I}\sum_{j=1}^{n_i} Z_{ij}q_{ij\boldsymbol{\lambda}} \qquad \text{where } q_{ij\boldsymbol{\lambda}} = \sum_{\ell=1}^{L}\lambda_\ell q_{ij\ell}. \end{aligned}$$

Of course, $T_\ell = T_{\boldsymbol{\lambda}}$ for $\boldsymbol{\lambda} = (0,0,\dots,0,1,0,\dots,0)^T$, where $\lambda_\ell = 1$ and $\lambda_{\ell'} = 0$ for $\ell' \neq \ell$; that is, T_ℓ is a statistic of the form $T_{\boldsymbol{\lambda}}$ for suitable $\boldsymbol{\lambda}$ and, from (3.1), $T_{\boldsymbol{\lambda}}$ is of the form $\sum_{i=1}^{I}\sum_{j=1}^{n_i} Z_{ij}q_{ij\boldsymbol{\lambda}}$ for suitable scores $q_{ij\boldsymbol{\lambda}}$, so the sensitivity calculations described in Section 2.3 apply directly to both T_ℓ and $T_{\boldsymbol{\lambda}}$, providing $\boldsymbol{\lambda}$ is selected a priori, without examining the data. In particular, $\{T_{\boldsymbol{\lambda}} - \mathrm{E}_{\Gamma,\mathbf{u}}(T_{\boldsymbol{\lambda}})\}/\sqrt{\mathrm{var}_{\Gamma,\mathbf{u}}(T_{\boldsymbol{\lambda}})} = \boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}})/\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{\mathbf{u}}\boldsymbol{\lambda}}$. Write $\Phi(\cdot)$ for the standard Normal cumulative distribution. To a close approximation for large I, the sensitivity analysis in Section 2.3 rejects H_0 in a one-sided test at level α if $\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}})/\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{\mathbf{u}}\boldsymbol{\lambda}} \geq \Phi^{-1}(1-\alpha)$ for (essentially) all $\mathbf{u} \in \mathcal{U}$ or, more precisely, for the one worst $\mathbf{u} \in \mathcal{U}$ constructed by the separable approximation.

Section 3.2 extends this reasoning to a $\boldsymbol{\lambda}$ selected by the investigator after examining the data.

3.2. *Sensitivity analysis for comparisons selected using the current data.* Proposition 1 speaks to the possibility that an investigator will become interested in a particular contrast $\boldsymbol{\lambda}$ after examining the data. Indeed, the investigator may try several values of $\boldsymbol{\lambda}$, or even all nonzero values of $\boldsymbol{\lambda}$, in an effort to find one that reports a high degree Γ of insensitivity to unmeasured biases, in the sense that, for this value of $\boldsymbol{\lambda}$, the deviate $\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}})/\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{\mathbf{u}}\boldsymbol{\lambda}}$ is large for all $\mathbf{u} \in \mathcal{U}$. There is, in Proposition 1, a price paid for multiple testing, for picking $\boldsymbol{\lambda}$ in light of the data, and later sections ask whether this price is worth paying.

Proposition 1 assumes H_0 is true for the purpose of testing it and assumes (2.3) is true for some specific but unknown $\mathbf{u}^*$ with a fixed $\gamma = \log(\Gamma)$ for the purpose of conducting one step in the sensitivity analysis. For any specific $\boldsymbol{\lambda}$, attention in (3.3) of Proposition 1 focuses on the $\mathbf{u} \in \mathcal{U}$ that minimizes $|\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})|/\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_\mathbf{u} \boldsymbol{\lambda}}$, thereby making H_0 difficult to reject; that is, one must reject H_0 with this $\mathbf{u}$ in order to report that rejection of H_0 is insensitive to a bias of magnitude Γ. Then (3.3) finds the nonzero $\boldsymbol{\lambda}$ that makes $\min_{\mathbf{u} \in \mathcal{U}} |\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})|/\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_\mathbf{u} \boldsymbol{\lambda}}$ as large as possible while controlling the probability of a false rejection at α. In other words, no matter how the investigator picks $\boldsymbol{\lambda}$, if the investigator rejects H_0 when $\min_{\mathbf{u} \in \mathcal{U}} |\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})|/\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_\mathbf{u} \boldsymbol{\lambda}} \geq \sqrt{c_\alpha}$, then the chance is at most α of rejecting H_0 when H_0 is true and the bias in (2.3) is at most $\Gamma = e^\gamma$. The critical constant c_α in Proposition 1 is discussed in Section 3.3.

PROPOSITION 1. *Suppose that the null hypothesis H_0 of the no treatment effect is true and treatment assignment $\mathbf{Z}$ has the distribution* (2.3) *for a specific $\gamma = \log(\Gamma) \geq 0$ and a specific but typically unknown $\mathbf{u}^* \in \mathcal{U}$. Let c_α be a constant such that*

$$\Pr\{(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*})^T \boldsymbol{\Sigma}_{\mathbf{u}^*}^{-1} (\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*}) \geq c_\alpha | \mathcal{F}, \mathcal{Z}\} \leq \alpha. \tag{3.2}$$

Then

$$\Pr\left\{\max_{\boldsymbol{\lambda} \neq \mathbf{0}} \min_{\mathbf{u} \in \mathcal{U}} \frac{|\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})|}{\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_\mathbf{u} \boldsymbol{\lambda}}} \geq \sqrt{c_\alpha} \middle| \mathcal{F}, \mathcal{Z}\right\} \leq \alpha. \tag{3.3}$$

PROOF. Using a standard result about extrema of quadratic forms [Rao (1973), page 60, 1f.1(i)],

$$\sqrt{(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})^T \boldsymbol{\Sigma}_\mathbf{u}^{-1} (\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})} = \max_{\boldsymbol{\lambda} \neq \mathbf{0}} \frac{|\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})|}{\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_\mathbf{u} \boldsymbol{\lambda}}} \qquad \text{for each } \mathbf{u} \in \mathcal{U}. \tag{3.4}$$

Then (3.3) follows from

$$\begin{aligned} \sqrt{(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*})^T \boldsymbol{\Sigma}_{\mathbf{u}^*}^{-1} (\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*})} &\geq \min_{\mathbf{u} \in \mathcal{U}} \sqrt{(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})^T \boldsymbol{\Sigma}_\mathbf{u}^{-1} (\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})} \\ &= \min_{\mathbf{u} \in \mathcal{U}} \max_{\boldsymbol{\lambda} \neq \mathbf{0}} \frac{|\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})|}{\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_\mathbf{u} \boldsymbol{\lambda}}} \\ &\geq \max_{\boldsymbol{\lambda} \neq \mathbf{0}} \min_{\mathbf{u} \in \mathcal{U}} \frac{|\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})|}{\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_\mathbf{u} \boldsymbol{\lambda}}}, \end{aligned} \tag{3.5}$$

where (3.5) uses a standard and straightforward inequality relating interchanging the order of max and min [e.g., Karlin (1992), Volume II, page 8, Lemma 1.3.1; Rosenbaum and Silber (2009b), Section 4.5]. □

3.3. *The critical constant c_α in Proposition 1.* As $I \to \infty$, under mild conditions, the smallest critical constant c_α such that (3.2) holds in Proposition 1 tends to the upper α critical value of the chi-square distribution on L degrees of freedom, say $\chi^2_{L,\alpha}$. The current section discusses sufficient conditions relevant to M-statistics and the stratified Wilcoxon statistic.

Before discussing these conditions, it is useful to indicate what this implies in the case of $L = 2$ test statistics, T_1 and T_2. With $L = 2$ test statistics, the value of $\sqrt{c_{0.05}}$ in (3.3) tends to 2.45 in testing all possible choices of $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$. For comparison, as $I \to \infty$, a two-sided α-level Bonferroni correction for two test statistics, T_1 and T_2, rejects if $|\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})|/\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_\mathbf{u}\boldsymbol{\lambda}} \geq \Phi^{-1}(1-\alpha/4)$ for either $\boldsymbol{\lambda} = (1,0)^T$ or $\boldsymbol{\lambda} = (0,1)^T$, where $\Phi^{-1}(1-0.05/4) = 2.24$. In other words, a price is paid for looking at all possible $\boldsymbol{\lambda}$ rather than just two values of $\boldsymbol{\lambda}$—the price is the move to 2.45 from 2.24—but it is not an extremely high price. Again, whether this price is worth paying is explored in a later section.

Under the assumptions of Proposition 1, $\mathrm{E}(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*}) = \mathbf{0}$ and $\mathrm{cov}(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*}) = \boldsymbol{\Sigma}_{\mathbf{u}^*}$. Using the familiar Cramér–Wold device, if $\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_\mathbf{u})/\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_\mathbf{u}\boldsymbol{\lambda}}$ converges in distribution to the standard Normal $\Phi(\cdot)$ for every $\boldsymbol{\lambda} \neq \mathbf{0}$, then $\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*}$ converges in distribution to the L-dimensional multivariate Normal with expectation $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{u}^*}$; see, for instance, Rao (1973), 2c.5(iv), page 128. Also, for each $\boldsymbol{\lambda} \neq \mathbf{0}$, the quantity $\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*})$ is the sum of I independent but not identically distributed random variables. In order to prove that $\Pr\{(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*})^T\boldsymbol{\Sigma}_{\mathbf{u}^*}^{-1}(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*}) \geq \chi^2_{L,\alpha}|\mathcal{F}, \mathcal{Z}\} \to \alpha$ as $I \to \infty$, it suffices to prove that the central limit theorem applies to $T_{\boldsymbol{\lambda}} = \boldsymbol{\lambda}^T\mathbf{T} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Z_{ij}q_{ij\boldsymbol{\lambda}}$ for each $\boldsymbol{\lambda} \neq \mathbf{0}$.

When does a central limit theorem apply to $T_{\boldsymbol{\lambda}} = \sum_{i=1}^{I}\sum_{j=1}^{n_i} Z_{ij}q_{ij\boldsymbol{\lambda}}$ for each $\boldsymbol{\lambda} \neq \mathbf{0}$? Recall that under (2.3) the $q_{ij\ell}$ are a sequence of constants and $T_{\boldsymbol{\lambda}}$ is random because treatment assignment Z_{ij} is random. It is convenient to assume that n_i is uniformly bounded, $n_i \leq \widetilde{n}$, say. If an M-statistic is used, it is additionally assumed that each quantile scale factor s_k in Section 3.1 converges to a positive limit. If each T_ℓ is a stratified Wilcoxon statistic or an M-statistic with bounded, continuous, monotone ψ-function, then the I independent random variables $\sum_{j=1}^{n_i} Z_{ij}q_{ij\boldsymbol{\lambda}}$ that are summed to produce $T_{\boldsymbol{\lambda}}$ are uniformly bounded. If $\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{\mathbf{u}^*}\boldsymbol{\lambda} \to \infty$ as $I \to \infty$, then the conditions of the Lindeberg central limit theorem are satisfied, and $\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}^*})/\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{\mathbf{u}^*}\boldsymbol{\lambda}}$ converges in distribution to the standard Normal. For M-statistics, one needs the double array version of this theorem [e.g., Billingsley (1979), Theorem 27.2, page 310] because s_k changes slightly as $I \to \infty$, altering all of the $q_{ij\boldsymbol{\lambda}}$.

The requirement that $\boldsymbol{\lambda}^T\boldsymbol{\Sigma}_{\mathbf{u}^*}\boldsymbol{\lambda} \to \infty$ is a requirement on the $q_{ij\ell}$ and it precludes various types of degeneracy taking hold as $I \to \infty$. For instance, if for all large i the responses of all n_i subjects in set i were equal, $R_{ijk} = R_{ij'k}$, $\forall i, j, j', k$, then $\sum_{j=1}^{n_i} Z_{ij}q_{ij\boldsymbol{\lambda}}$ would be constant for large i and the central limit theorem would not apply; however, $\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_{\mathbf{u}^*} \boldsymbol{\lambda} \to \infty$ precludes this. Because $\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_{\mathbf{u}^*} \boldsymbol{\lambda}$ must tend to ∞ for all $\boldsymbol{\lambda} \neq \mathbf{0}$, the condition also precludes a limit for $\boldsymbol{\Sigma}_{\mathbf{u}^*}$ that is not positive definite.

3.4. *Closed testing of subhypotheses.* Proposition 1 provides a basis for testing the null hypothesis H_0 of no effect on all K outcomes. If this null hypothesis is rejected, then subhypotheses may be examined using the closed testing method of Marcus, Peritz and Gabriel (1976), as implemented for multiple outcomes by Lehmacher, Wassmer and Reitmeir (1991). To use closed testing, take $L = K$ with test statistic T_k computed from response k. Stated informally, if H_0 is rejected for all K outcomes, then K analogous hypotheses are tested concerning no effect on $K - 1$ outcomes, now with a slightly more generous constant c_α, and so on, terminating a branch of testing when an acceptance occurs, possibly testing individual outcomes; see Lehmacher, Wassmer and Reitmeir (1991) for the specifics of converting one multivariate test into a closed testing procedure. Closed testing has attractive properties when used in sensitivity analyses; see Rosenbaum and Silber (2009b).

4. Design sensitivity and power with fixed and discovered contrasts.

4.1. *Design sensitivity with fixed* $\boldsymbol{\lambda}$. If the treatment did have an effect and if there were actually no bias from unmeasured covariates, then we could not recognize this situation had occurred from the observable data. Not knowing that we are in this *favorable situation* with an effect and no bias, the best we could hope to report is that rejection of the null hypothesis H_0 of no effect is insensitive to small and moderate biases Γ. The power of an α-level sensitivity analysis is the probability that this hope will be realized, that is, the probability of rejection of H_0 at level α using a particular test while allowing for a bias of Γ in a particular design or sampling situation with a treatment effect and no unmeasured bias; see Rosenbaum (2004). Under mild conditions, there is a value $\widetilde{\Gamma}$ called the design sensitivity such that, for every $\alpha > 0$, the power of the sensitivity analysis tends to 1 as $I \to \infty$ for $\Gamma < \widetilde{\Gamma}$ and to 0 for $\Gamma > \widetilde{\Gamma}$, so $\widetilde{\Gamma}$ is a concise indicator of large sample power [Rosenbaum (2004); (2010), Part III; (2013, 2014)]. Moreover, $\widetilde{\Gamma}$ is closely connected with the Bahadur efficiency of the sensitivity analysis; see Rosenbaum (2015). In short, computing the power of a sensitivity analysis or the design sensitivity means assuming that: (i) we are in the favorable situation with data generated by a specific stochastic model with a treatment effect and no unmeasured biases, (ii) we are, as we would be in practice, ignorant of the fact that we are in the favorable situation, so we conduct sensitivity analyses with various values of Γ, (iii) we evaluate the stochastic performance of these sensitivity analyses in this favorable situation, in particular, considering the probability that we reject H_0 at level α, allowing for a bias of magnitude Γ.

Because for fixed $\boldsymbol{\lambda}$ the statistic $T_{\boldsymbol{\lambda}} = \sum_{i=1}^{I} \sum_{j=1}^{n_i} Z_{ij} q_{ij\boldsymbol{\lambda}}$ is essentially a univariate statistic, the calculation of the design sensitivity of $T_{\boldsymbol{\lambda}}$ for a fixed $\boldsymbol{\lambda}$ closely parallels existing results. In general, the numerical calculation of the design sensitivity $\widetilde{\Gamma}$ for matched sets with more than one control, $n_i \geq 3$, is not difficult, but it does not produce a simple formula; see Rosenbaum (2004, 2013, 2014) for such calculations. To exhibit a formula for $\widetilde{\Gamma}$, consider the case of matched pairs, $n_i = 2$, using an M-statistic with an odd, continuous, bounded, monotone increasing ψ-function that is not identically zero, for instance, $\psi_{\text{hu}}(y)$, and suppose that the favorable situation consists of I matched pairs of I independent and identically distributed (i.i.d.) observations from a sampling situation with a treatment effect and no unmeasured biases so that $\Pr(\mathbf{Z} = \mathbf{z} | \mathcal{F}, \mathcal{Z}) = |\mathcal{Z}|^{-1}$ for each $\mathbf{z} \in \mathcal{Z}$. In this case, write Y_{ik} for the treated-minus-control pair difference in response k in pair i, so $Y_{ik} = (Z_{i1} - Z_{i2})(R_{i1k} - R_{i2k}) = (Z_{i1} - Z_{i2}) D_{i12k}$, and s_k is a quantile, typically the median, of the $|Y_{ik}|$, and write ω_k for the corresponding population quantile. Because $\psi(\cdot)$ is odd, the case of matched pairs simplifies, with $\sum_{j=1}^{2} Z_{ij} \sum_{j'=1}^{2} \psi(D_{ijj'k}/s_k) = \psi(Y_{ik}/s_k)$, so that $T_k = \sum_{i=1}^{I} \psi(Y_{ik}/s_k)$ and $T_{\boldsymbol{\lambda}} = \sum_{i=1}^{I} \sum_{k=1}^{K} \lambda_k \psi(Y_{ik}/s_k)$. Then the I quantities $\sum_{k=1}^{K} \lambda_k \psi(Y_{ik}/\omega_k)$ are i.i.d., so write $\theta_{\boldsymbol{\lambda}} = \mathrm{E}\{\sum_{k=1}^{K} \lambda_k \psi(Y_{ik}/\omega_k)\}$ and $\eta_{\boldsymbol{\lambda}} = \mathrm{E}\{|\sum_{k=1}^{K} \lambda_k \psi(Y_{ik}/\omega_k)|\}$, where both expectations exist because $\psi(\cdot)$ is bounded, and of course $\eta_{\boldsymbol{\lambda}} \geq \theta_{\boldsymbol{\lambda}}$ with strict inequality unless $\sum_{k=1}^{K} \lambda_k \psi(Y_{ik}/\omega_k)$ is non-negative with probability 1.

PROPOSITION 2. *In the case of i.i.d. matched pairs*, $n_i = 2$, *for fixed* $\boldsymbol{\lambda}$, *the design sensitivity* $\widetilde{\Gamma}_{\boldsymbol{\lambda}}$ *of* $T_{\boldsymbol{\lambda}} = \sum_{i=1}^{I} \sum_{k=1}^{K} \lambda_k \psi(Y_{ik}/s_k)$ *is*

$$\widetilde{\Gamma}_{\boldsymbol{\lambda}} = \frac{\eta_{\boldsymbol{\lambda}} + \theta_{\boldsymbol{\lambda}}}{\eta_{\boldsymbol{\lambda}} - \theta_{\boldsymbol{\lambda}}} \qquad \textit{if } \eta_{\boldsymbol{\lambda}} > \theta_{\boldsymbol{\lambda}} \textit{ and is } \widetilde{\Gamma} = \infty \textit{ otherwise.} \tag{4.1}$$

The proof of Proposition 2 is almost the same as the proof of Corollary 1 in Rosenbaum (2013) and is omitted. The idea of the proof is that $T_{\boldsymbol{\lambda}}/I$ converges in probability to $\theta_{\boldsymbol{\lambda}}$, whereas, in the paired case, $\max_{\mathbf{u} \in \mathcal{U}} \boldsymbol{\lambda}^T \boldsymbol{\mu}_{\mathbf{u}}$ converges in probability to $\eta_{\boldsymbol{\lambda}}(\Gamma - 1)/(\Gamma + 1)$; then $\widetilde{\Gamma}$ in (4.1) is obtained by equating these limits and solving for Γ.

4.2. *Numerical evaluation of the optimal design sensitivity.* Table 2 evaluates the design sensitivity $\widetilde{\Gamma}$ in (4.1) in the case of matched pairs, $n_i = 2$, with i.i.d. bivariate outcomes, $K = 2$, having either a bivariate Normal distribution or a bivariate t-distribution with 5 degrees of freedom. In Table 2, the treated-minus-control matched pair differences Y_{i1} for the first outcome always have an expectation that is half its standard deviation, whereas this effect size ranges from 0 to 0.5 for the second outcome, Y_{i2}. Table 2 reports the design sensitivity for the first outcome alone, $\widetilde{\Gamma}_1$, for the second outcome alone, $\widetilde{\Gamma}_2$, and for an optimal combination of the two outcomes, $\widetilde{\Gamma}_{\boldsymbol{\lambda}}$. Neither $\widetilde{\Gamma}_1$ nor $\widetilde{\Gamma}_2$ depends upon the correlation parameter ρ, but the optimal $\widetilde{\Gamma}_{\boldsymbol{\lambda}}$ does depend on ρ.

TABLE 2

Design sensitivities with bivariate Normal errors or bivariate t errors with five degrees of freedom (t_5) errors. Effect sizes τ_k are the expected treated-minus-control pair differences in units of the standard deviation, with $\tau_1 = 0.5$ and τ_2 varying

τ_2	ρ	$\widetilde{\Gamma}_1$	$\widetilde{\Gamma}_2$	$\widetilde{\Gamma}_\lambda$	**Optimal λ_2**
		Normal distribution, ψ_{hu}			
0.00	0.0	3.4	1.0	3.4	0.00
0.00	0.5	3.4	1.0	4.1	−0.48
0.25	0.0	3.4	1.8	4.0	0.49
0.25	0.5	3.4	1.8	3.4	0.01
0.50	0.0	3.4	3.4	5.8	1.00
0.50	0.5	3.4	3.4	4.1	1.00
		Normal distribution, ψ_{in}			
0.00	0.0	4.2	1.0	4.2	0.00
0.00	0.5	4.2	1.0	4.5	−0.29
0.25	0.0	4.2	2.0	4.4	0.25
0.25	0.5	4.2	2.0	4.2	0.00
0.50	0.0	4.2	4.2	6.2	1.00
0.50	0.5	4.2	4.2	4.6	1.00
		t_5 distribution, ψ_{hu}			
0.00	0.0	3.8	1.0	3.8	0.00
0.00	0.5	3.8	1.0	4.7	−0.47
0.25	0.0	3.8	2.0	4.5	0.50
0.25	0.5	3.8	2.0	3.8	0.03
0.50	0.0	3.8	3.8	6.8	1.00
0.50	0.5	3.8	3.8	4.7	1.00
		t_5 distribution, ψ_{in}			
0.00	0.0	4.4	1.0	4.4	0.00
0.00	0.5	4.4	1.0	4.8	−0.29
0.25	0.0	4.4	2.1	4.7	0.30
0.25	0.5	4.4	2.1	4.4	0.00
0.50	0.0	4.4	4.4	6.8	1.00
0.50	0.5	4.4	4.4	5.0	1.00

The test statistic $T_\lambda = \sum_{k=1}^{2} \lambda_k \sum_{i=1}^{I} \psi(Y_{ik}/s_k) = \lambda_1 T_1 + \lambda_2 T_2$ is a weighted combination of two M-statistics, where s_k is the median $|Y_{ik}|$ and the ψ-function is either Huber's $\psi_{\mathrm{hu}}(y) = \mathrm{sign}(y)\min(|y|, \kappa)$ or a version that performs inner trimming, $\psi_{\mathrm{in}}(y) = \mathrm{sign}(y)\{\kappa/(\kappa - \iota)\}\max\{0, \min(|y|, \kappa) - \iota\}$, with $0 \le \iota < \kappa$, so $|\psi_{\mathrm{in}}(y)|$ is 0 for $|y| \in [0, \iota]$, is κ for $|y| \in [\kappa, \infty)$ and rises linearly from 0 to κ on (ι, κ). Inner trimming has been shown to increase the power of a sensitivity analysis and to increase design sensitivity; see Rosenbaum (2013). In Table 2, $\iota = 0.5$ and $\kappa = 2.5$. The optimal $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ is found by setting $\lambda_1 = 1$ and numerically optimizing (4.1) over λ_2. The final column of Table 2 reports the optimal λ_2.

Some specifics follow. Let $(\varepsilon_{i1}, \varepsilon_{i2})$, $i = 1, \dots, I$, be independent bivariate Normal random vectors with zero expectations, unit variances and correlation ρ. For bivariate Normal errors, $(Y_{i1}, Y_{i2}) = (\tau_1, \tau_2) + (\varepsilon_{i1}, \varepsilon_{i2})$. For bivariate t errors with 5 degrees of freedom, $(Y_{i1}, Y_{i2}) = \sqrt{5/3}(\tau_1, \tau_2) + (\varepsilon_{i1}, \varepsilon_{i2})/\sqrt{\chi^2_{i,5}/5}$, where $\chi^2_{i,5}$ are independent chi-square random variables with 5 degrees of freedom, independent of the $(\varepsilon_{i1}, \varepsilon_{i2})$'s, so $\mathrm{var}(Y_{ik}) = 5/3$ and $\mathrm{E}(Y_{ik})/\sqrt{\mathrm{var}(Y_{ik})} = \tau_k$. In other words, for both the Normal and t-distributions, τ_k is the expected treated-minus-control pair difference in units of the standard deviation, $\tau_k = \mathrm{E}(Y_{ik})/\sqrt{\mathrm{var}(Y_{ik})}$. As is always true with the bivariate t-distribution, ρ is the correlation of the Normal $(\varepsilon_{i1}, \varepsilon_{i2})$, but not of (Y_{i1}, Y_{i2}) for the t-distribution. Numerical calculations used the `mvtnorm` package in `R`; see Genz and Bretz (2009).

In Table 2, consider, first, $\widetilde{\Gamma}_1$ and $\widetilde{\Gamma}_2$ for the two outcomes analyzed separately. When the second outcome is unaffected, $\tau_2 = 0$, the design sensitivity is, of course, $\widetilde{\Gamma}_2 = 1$, whereas when $\tau_1 = \tau_2 = 0.5$, the separate design sensitivities are equal, $\widetilde{\Gamma}_1 = \widetilde{\Gamma}_2$. As in Rosenbaum (2013), inner trimming $\psi_{\mathrm{in}}(\cdot)$ yields somewhat higher design sensitivities than $\psi_{\mathrm{hu}}(\cdot)$ in the sampling situations in Table 2.

Consider now the design sensitivity $\widetilde{\Gamma}_{\boldsymbol{\lambda}}$ for $T_{\boldsymbol{\lambda}}$ when $\boldsymbol{\lambda}$ is chosen to maximize $\widetilde{\Gamma}_{\boldsymbol{\lambda}}$. If $\rho = 0$ and $\tau_2 = 0$, the optimal weight ignores Y_{i2} with $\lambda_2 = 0$ and $\widetilde{\Gamma}_{\boldsymbol{\lambda}} = \widetilde{\Gamma}_1$. If two outcomes are equally affected by the treatment, $\tau_1 = \tau_2 = 0.5$, then the optimal weights are equal, $\lambda_1 = \lambda_2 = 1$, and $\widetilde{\Gamma}_{\boldsymbol{\lambda}} > \max(\widetilde{\Gamma}_1, \widetilde{\Gamma}_2)$, with the difference $\widetilde{\Gamma}_{\boldsymbol{\lambda}} - \max(\widetilde{\Gamma}_1, \widetilde{\Gamma}_2)$ being quite large when $\rho = 0$ and not small for $\rho = 0.5$. For uncorrelated outcomes where the second outcome is less affected than the first, $\rho = 0$ and $\tau_2 = 0.25$, the optimal weight has $0 < \lambda_2 < 1$ and a smaller increase in design sensitivity, $\widetilde{\Gamma}_{\boldsymbol{\lambda}} - \max(\widetilde{\Gamma}_1, \widetilde{\Gamma}_2)$. A particularly interesting case with $\widetilde{\Gamma}_{\boldsymbol{\lambda}} > \max(\widetilde{\Gamma}_1, \widetilde{\Gamma}_2)$ has $\rho = 0.5$ and $\tau_2 = 0$, so the two outcomes are correlated, but the second outcome is unaffected by the treatment—in this case, Y_{i2} has sometimes been called a "control outcome"; see McKillip (1992), Weiss (2002) and Rosenbaum (2010), Section 5.2.4. With $\rho = 0.5$ and $\tau_2 = 0$, the optimal λ_2 is negative, so $T_{\boldsymbol{\lambda}}$ is large when the affected T_1 is larger than the unaffected T_2. Between these two situations, with a correlated but smaller effect, $\rho = 0.5$ and $\tau_2 = 0.25$, the optimal λ_2 is close to 0 and $\widetilde{\Gamma}_{\boldsymbol{\lambda}}$ is negligibly different from $\widetilde{\Gamma}_1$.

In many of the sampling situations in Table 2, in sufficiently large samples, the optimal $T_{\boldsymbol{\lambda}}$ will report greater insensitivity to unmeasured biases than both of its components, T_1 and T_2. For instance, in the four situations in Table 2 with $\tau_1 = \tau_2 = 0.5$ and $\rho = 0$, the power of a sensitivity analysis performed with $\Gamma = 5$ is tending to 0 as $I \to \infty$ for both T_1 and T_2, but the power is tending to 1 for the optimal $T_{\boldsymbol{\lambda}}$. Moreover, with $\tau_1 = \tau_2 = 0.5$ and $\rho = 0$ at $\Gamma = 5$, the chance that T_1 or T_2 rejects H_0 but $T_{\boldsymbol{\lambda}}$ does not is declining to 0 at an exponential rate with increasing I, so this event is improbable even for moderate sample sizes; see Rosenbaum (2015).

4.3. *Design sensitivity when the optimal* $\boldsymbol{\lambda}$ *is unknown.* Suppose that the treatment has an effect and there is no unmeasured bias, and suppose $\widetilde{\boldsymbol{\lambda}}$ is any $\boldsymbol{\lambda}$ such that $\widetilde{\Gamma}_{\widetilde{\boldsymbol{\lambda}}} = \max_{\boldsymbol{\lambda}\neq\mathbf{0}} \widetilde{\Gamma}_{\boldsymbol{\lambda}}$. For the test statistic $T_{\widetilde{\boldsymbol{\lambda}}}$ with this best $\widetilde{\boldsymbol{\lambda}}$, define the deviate $A_{\Gamma} = \min_{\mathbf{u}\in\mathcal{U}} |\widetilde{\boldsymbol{\lambda}}^T(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}}^{(\Gamma)})| / \sqrt{\widetilde{\boldsymbol{\lambda}}^T \boldsymbol{\Sigma}_{\mathbf{u}}^{(\Gamma)} \widetilde{\boldsymbol{\lambda}}}$. By the definition of the design sensitivity, if $\Gamma < \widetilde{\Gamma}_{\widetilde{\boldsymbol{\lambda}}}$, then the power $\Pr(A_{\Gamma} \geq a | \mathcal{F}, \mathcal{Z}) \to 1$ as $I \to \infty$ for each a. Because $\widetilde{\boldsymbol{\lambda}}$ depends upon the unknown sampling distribution of the responses, R_{ijk}, it is not possible to test H_0 using this optimal $T_{\widetilde{\boldsymbol{\lambda}}}$. For instance, in Table 2, determining the optimal $\widetilde{\boldsymbol{\lambda}}$ required knowledge of the distribution of (Y_{i1}, Y_{i2}).

Instead, define

$$A_{\Gamma}^* = \max_{\boldsymbol{\lambda}\neq\mathbf{0}} \min_{\mathbf{u}\in\mathcal{U}} \frac{|\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}}^{(\Gamma)})|}{\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_{\mathbf{u}}^{(\Gamma)} \boldsymbol{\lambda}}}.$$

By Proposition 1, if H_0 were true and the bias in treatment assignment was at most Γ, then a test that rejected H_0 when $A_{\Gamma}^* \geq \sqrt{c_\alpha}$ would falsely reject H_0 with probability at most α. Moreover, by definition, $A_{\Gamma}^* \geq A_{\Gamma}$. If $\Gamma < \widetilde{\Gamma}_{\widetilde{\boldsymbol{\lambda}}}$, then it follows that the power $\Pr(A_{\Gamma}^* \geq \sqrt{c_\alpha} | \mathcal{F}, \mathcal{Z})$ of the test that rejects H_0 when $A_{\Gamma}^* \geq \sqrt{c_\alpha}$ tends to 1 as $I \to \infty$. So the feasible test using A_{Γ}^* achieves the same design sensitivity as the test using $T_{\widetilde{\boldsymbol{\lambda}}}$ (or, equivalently, A_{Γ}), which is not feasible because it requires knowledge of the optimal $\widetilde{\boldsymbol{\lambda}}$.

4.4. *Simulated power of sensitivity analyses.* Where Sections 4.1–4.3 consider the limit as the number I of matched sets increased, $I \to \infty$, the current section evaluates the power in finite samples of the procedure in Proposition 1. Table 3 considers twelve of the 24 sampling situations from Table 2, that is, matched pairs, $n_i = 2$, bivariate observations, $K = 2$, from either the bivariate Normal distribution or the bivariate t-distribution with 5 degrees of freedom. The effect size τ_k for coordinate k, $k = 1, 2$, is in units of the standard deviation of a matched pair difference, Y_{ik}, so $\tau_k = \mathrm{E}(Y_{ik})/\sqrt{\mathrm{var}(Y_{ik})}$. In all cases, the effect size for the first coordinate is $\tau_1 = 0.5$, but τ_2 is either 0 or 0.5. In Table 2, the correlation parameter for the underlying Normal distribution is either $\rho = 0$ or $\rho = 0.5$. If $\tau_2 = 0$ and $\rho = 0$, then Y_{i2} is irrelevant and an optimal weighting ignores Y_{i2}. If $\tau_2 = 0.5$ and $\rho = 0$, then Y_{i2} is as informative as Y_{i1} and an optimal weighting gives the two coordinates equal weights. If $\tau_2 = 0$ and $\rho = 0.5$, then Y_{i2} is a "control outcome" and an optimal weighting gives the second coordinate a negative weight.

Table 3 reports the power of a 0.05-level sensitivity analysis when conducted with $\Gamma = 3$ in the favorable situation with a treatment effect and no unmeasured bias. Specifically, in 10,000 samples of size $I = 500$ pairs, Table 3 reports the proportion of upper bounds on the P-value that were 0.05 or less, thereby saying that a bias of $\Gamma = 3$ is too small to explain the observed association between treatment and outcome. With 10,000 replicates, the standard error of a simulated power is at most $\sqrt{0.25/10{,}000} = 0.005$.

Table 3 reports the power of five procedures, of which only two are practical procedures, the others serving as benchmarks for comparison. Three of the procedures at the far right [columns (iv)–(vi)] do not correct for multiple testing, so

TABLE 3

Simulated power of 0.05*-level sensitivity analyses at* $\Gamma = 3$ *in* $I = 500$ *pairs*, $n_i = 2$, *with bivariate Normal errors* (*N*) *or bivariate t errors with five degrees of freedom* (t_5). *Effect sizes* τ_k *are the expected treated-minus-control pair differences in units of the standard deviation, with* $\tau_1 = 0.5$ *and* τ_2 *varying. Each sampling situation is replicated* 10,000 *times. Uncorrected tests are based on one-sided P-values with no correction for multiple testing. Corrected tests correct for multiple testing. In each sampling situation, the highest corrected power is in* ***bold***

Sampling distribution			**Corrected tests**			**Uncorrected tests**		
Column number			**(i)**	**(ii)**	**(iii)**	**(iv)**	**(v)**	**(vi)**
	τ_2	ρ	**Bonferroni**	**Maximum**	**Optimal**	**First**	**Second**	**Optimal**
				Normal distribution, ψ_{hu}				
N	0.0	0.0	**0.08**	0.06	0.05	0.23	0.00	0.23
N	0.0	0.5	0.07	**0.46**	0.43	0.22	0.00	0.79
N	0.5	0.0	0.14	**1.00**	0.99	0.22	0.22	1.00
				Normal distribution, ψ_{in}				
N	0.0	0.0	**0.43**	0.34	0.34	0.70	0.00	0.70
N	0.0	0.5	0.44	**0.62**	0.59	0.70	0.00	0.89
N	0.5	0.0	0.69	**1.00**	1.00	0.71	0.70	1.00
				t_5 *distribution*, ψ_{hu}				
t_5	0.0	0.0	**0.32**	0.26	0.24	0.58	0.00	0.58
t_5	0.0	0.5	0.31	**0.80**	0.78	0.57	0.00	0.95
t_5	0.5	0.0	0.52	**1.00**	1.00	0.56	0.57	1.00
				t_5 *distribution*, ψ_{in}				
t_5	0.0	0.0	**0.59**	0.50	0.50	0.81	0.00	0.81
t_5	0.0	0.5	0.58	**0.76**	0.73	0.81	0.00	0.94
t_5	0.5	0.0	0.82	**1.00**	1.00	0.81	0.81	1.00

these procedures would be appropriate only if selected in advance, without examining the data. Two of the procedures labeled "optimal" [columns (iii) and (vi)] use the optimal $\boldsymbol{\lambda}$ for large I from Table 2, but this optimal $\boldsymbol{\lambda}$ is derived from the true sampling distribution and is unknown to the investigator. The "corrected maximum" test [column (ii)] is based on Proposition 1, maximizing over $\boldsymbol{\lambda} \neq \mathbf{0}$ the minimum deviate over $\mathbf{u} \in \mathcal{U}$, with $\sqrt{c_{0.05}} = 2.45$ as in Section 3.3. The Bonferroni procedure [column (i)] tests both coordinates of (Y_{i1}, Y_{i2}) in both tails so the absolute value of a standardized deviate must exceed 2.24 to produce a P-value bound of 0.05 or less, as in Section 3.3. The uncorrected tests are all one-sided, so they assume the investigator knows the direction of the effect, and H_0 is rejected at the 0.05 level if the deviate exceeds $1.645 = \Phi^{-1}(0.95)$. In contrast, the corrected optimal procedure [column (iii)] uses optimal weights but with the critical value $\sqrt{c_{0.05}} = 2.45$. Keep in mind that, because it capitalizes on chance in its choice of $\boldsymbol{\lambda}$, the maximum deviate in (3.3) is almost always slightly larger than the deviate that uses the optimal $\boldsymbol{\lambda}$ from Table 2, so the power in column (ii) is slightly higher than in column (iii). In other words, the critical constant $\sqrt{c_{0.05}} = 2.45$ corrects for capitalizing on chance, and column (ii) does capitalize on chance, but column (iii) does not because it uses the unknown optimal $\boldsymbol{\lambda}$.

Obviously, it is best to know what you do not know, the optimal $\boldsymbol{\lambda}$ and the direction of the effect, so the unattainable power in column (vi) is the highest in every row of Table 2. Importantly, this optimal power in column (vi) is strictly higher than the maximum power in columns (iv) and (v) except when Y_{i2} is irrelevant ($\tau_2 = 0$ and $\rho = 0$), so there is something to be gained by looking at both outcomes in combination. If the same optimal weights and deviate are used with the larger critical constant of $\sqrt{c_{0.05}} = 2.45$ rather than 1.645, then of course the power drops, as seen by comparing columns (vi) to column (iii); that is, the comparison of columns (iii) and (vi) shows the pure effect of changing the critical constant to 2.45 from 1.645 with the same standardized deviate.

Comparing the two practical procedures in columns (i) and (ii), the Bonferroni procedure has somewhat higher power when Y_{i2} is irrelevant because then the optimal $\boldsymbol{\lambda}$ ignores Y_{i2}, but in the other two cases the Bonferroni procedure has substantially lower power. Columns (ii) and (iii) are quite similar. In brief, the power in column (ii) for the procedure in Proposition 1 is lower than the optimal power in column (vi) mostly because it corrects the critical value for discovering a good $\boldsymbol{\lambda}$ using the data at hand, and the power is sometimes much higher than for the Bonferroni procedure in column (i) mostly because it is similar to using the optimal $\boldsymbol{\lambda}$ in column (iii).

Many modern observational studies are based on administrative or survey data and have sample sizes orders of magnitude larger than $I = 500$. For such studies, Table 2 provides better guidance than Table 3.

5. Sensitivity analysis in the periodontal data.

5.1. *Sensitivity analysis using a standard test statistic.* The sensitivity analysis for the periodontal data in Section 1.1 will first use an M-statistic with Huber's $\psi_{\mathrm{hu}}(\cdot)$ with $\kappa = 2.5$, the default ψ-function in the `senmv` function in the `sensitivitymv` package (version 1.3) in `R`. In Section 5.2, the sensitivity analyses will be repeated with the inner trimmed $\psi_{\mathrm{in}}(\cdot)$. As $K = L = 2$, the critical value $c_{0.05}$ from the chi-square distribution on 2 degrees of freedom is $c_{0.05} = 5.9915$ or $\sqrt{c_{0.05}} = 2.4477$, and the argument in Section 3.3 shows this critical constant is appropriate as $I \to \infty$ with these ψ-functions. Using $\psi_{\mathrm{hu}}(\cdot)$, the test in Proposition 1 just barely rejects H_0 at the 0.05 level for $\Gamma = 2.2$ because, with this Γ, the largest deviate is

$$(5.1) \qquad \max_{\boldsymbol{\lambda} \neq \mathbf{0}} \min_{\mathbf{u} \in \mathcal{U}} \frac{|\boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}}^{(\Gamma)})|}{\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_{\mathbf{u}}^{(\Gamma)} \boldsymbol{\lambda}}} = 2.4513 > 2.4477 = \sqrt{c_{0.05}},$$

which is attained with $\boldsymbol{\lambda} = (0.714, 0.286)^T$. In words, the least sensitive combination of two M-statistics gave positive weight to both lower and upper teeth, but gave more than two times as much weight to the lower teeth. The bivariate pattern seen in Figure 2 cannot quite be explained by a bias of $\Gamma = 2.2$, but every $\boldsymbol{\lambda}$ produces a minimum deviate less than $\sqrt{c_{0.05}}$ for $\Gamma = 2.3$.

An aid to interpreting the value of Γ is a device called amplification; see Rosenbaum and Silber (2009a) and the `amplify` function in the `sensitivitymv` package, version 1.3, in `R`. A single value of $\Gamma \geq 1$ is equivalent to a curve of values of two parameters, $\Gamma = (\Delta\Lambda + 1)/(\Delta + \Lambda)$, where an unobserved covariate u increases the odds of treatment by a factor of Λ and increases the odds of a positive matched pair difference in responses under control, $r_{Ci1} - r_{Ci2}$, by a factor of Δ. For instance, a bias of $\Gamma = 2.2$ corresponds with an unobserved covariate that increases the odds of smoking by more than 4-fold and increases the odds of a positive pair difference in periodontal disease by more than 4-fold because $\Gamma = 2.2 > 2.125 = (4 \times 4 + 1)/(4 + 4)$. Similarly, $\Gamma = 2.2 = (3 \times 7 + 1)/(3 + 7)$, so the curve $\Gamma = (\Delta\Lambda + 1)/(\Delta + \Lambda)$ includes $(\Delta, \Lambda) = (3, 7)$ and $(\Delta, \Lambda) = (7, 3)$; that is, $\Gamma = 2.2$ corresponds with an unobserved u that triples the odds of smoking and increases the odds of greater periodontal disease by 7-fold, but it also corresponds with an unobserved covariate that increases the odds of smoking by 7-fold and increases the odds of greater periodontal disease by 3-fold. The correspondence is that a sensitivity analysis for any (Δ, Λ) on the curve $\Gamma = (\Delta\Lambda + 1)/(\Delta + \Lambda)$ gives exactly the same results as the one sensitivity using Γ; see Rosenbaum and Silber (2009a) for technical details. The amplification helps to understand the magnitude of Γ in terms of its impact in a simple situation, namely, matched pairs. To explain the association in Figure 2 as something other than an effect caused by smoking, the bias from an unobserved covariate would need to be larger than this.

At $\Gamma = 2.2$, the minimum deviate, $\min_{\mathbf{u} \in \mathcal{U}} \boldsymbol{\lambda}^T(\mathbf{T} - \boldsymbol{\mu}_{\mathbf{u}}^{(\Gamma)})/\sqrt{\boldsymbol{\lambda}^T \boldsymbol{\Sigma}_{\mathbf{u}}^{(\Gamma)} \boldsymbol{\lambda}}$, is 2.155 for lower teeth, $\boldsymbol{\lambda} = (1, 0)^T$, yielding a two-sided Bonferroni corrected P-value of $4\Phi(-2.15) = 0.062$, whereas for upper teeth, $\boldsymbol{\lambda} = (0, 1)^T$, the minimum deviate is 0.599 with Bonferroni corrected P-value of 1. So at $\Gamma = 2.2$, the Bonferroni method would test its two null hypotheses and fail to reject both of them, while Proposition 1 would reject the bivariate H_0 at the family-wise 0.05 level having tested infinitely many hypotheses. In words, looking at upper teeth and lower teeth separately would lead us to conclude that a bias of $\Gamma = 2.2$ could explain Figure 2 as something other than an effect caused by smoking, whereas (5.1) disagrees, saying a bias of $\Gamma = 2.2$ is too small to explain the ostensible effect of smoking. Using Proposition 1 at $\Gamma = 2.2$ to produce adjusted P-values from the chi-square distribution with 2 degrees of freedom, the adjusted P-value for $\boldsymbol{\lambda} = (0.714, 0.286)^T$ is 0.0496, for $\boldsymbol{\lambda} = (1, 0)^T$ is 0.0982, for $\boldsymbol{\lambda} = (0, 1)^T$ is 0.836 and for $\boldsymbol{\lambda} = (0.5, 0.5)^T$ is 0.0732.

Consider, now, the closed testing procedure in Section 3.4. In closed testing, if the bivariate hypothesis is rejected, the univariate hypotheses are tested as uncorrected two-sided tests. Closed testing may be implemented either using Proposition 1 or using the Bonferroni inequality, and in the latter case it becomes Holm's

(1979) procedure. At $\Gamma = 2.2$, the Bonferroni/Holm closed testing procedure does not reject any hypothesis. At $\Gamma = 2.2$, the contrast $\boldsymbol{\lambda} = (0.714, 0.286)^T$ in (5.1) barely rejects H_0, and closed testing then compares the deviate 2.155 for lower teeth, $\boldsymbol{\lambda} = (1, 0)^T$, to 1.96, the two-sided 0.05 critical value from the Normal distribution, so closed testing rejects the hypothesis of no effect on lower teeth. Not only did Proposition 1 report greater insensitivity to unmeasured bias for the bivariate outcome, but it also reported greater insensitivity to unmeasured bias for a single outcome, lower teeth, when Proposition 1 formed the basis for closed testing.

If closed testing is applied at $\Gamma = 1.8$, then Scheffé projections reject the two hypotheses of no effect for upper and no effect for lower teeth. In other words, there is evidence of an effect on lower teeth and on upper teeth, but a smaller bias Γ could explain the ostensible effect on upper teeth.

Consider point estimates and 95% confidence intervals for an additive or shift effect for each outcome separately with a bias of at most $\Gamma = 1.5$. The shift effect is the typical increase in the number of measurements with either a loss of attachment or a pocket depth of ≥ 4 mm. A bias of $\Gamma = 1.5$ is equivalent to an unobserved covariate that doubles the odds of treatment and increases by 4-fold the odds of a positive pair difference in responses; see Rosenbaum and Silber (2009a) and the `amplify` function in `sensitivitymv`. When $\Gamma > 1$, there is not a single point estimate, but rather an interval of point estimates, and at $\Gamma = 1.5$ that interval is $[3.34, 8.72]$ for lower teeth and $[1.51, 5.76]$ for upper teeth. Of course, the corresponding 2-sided 95% confidence intervals are wider because they allow for both sampling variability and a bias of $\Gamma = 1.5$; they are $[2.01, 10.64]$ for lower teeth and $[0.63, 7.43]$ for upper teeth. The calculations in this paragraph used the `senmwCI` function in the `sensitivitymw` package in `R` using the method in Rosenbaum (2007), and, unlike the tests above, the confidence intervals are not simultaneous intervals.

5.2. *Sensitivity analysis using a statistic that emphasizes larger effects.* Tables 2 and 3 and results in the literature anticipate higher power in a sensitivity analysis and a larger design sensitivity if $\psi_{\text{in}}(\cdot)$ is used in place of $\psi_{\text{hu}}(\cdot)$. This anticipation is correct for the periodontal data in Figure 2. Using $\psi_{\text{in}}(\cdot)$ with $\iota = 0.5$ and $\kappa = 2.5$, the statistic in (5.1) is 2.460 at $\Gamma = 2.37$, leading to rejection at the 0.05 level of the hypothesis of no effect on the bivariate outcome, and closed testing goes on to reject the hypothesis of no effect on lower teeth. At $\Gamma = 2.37$ using $\psi_{\text{in}}(\cdot)$, the Bonferroni/Holm procedure also rejects the null hypothesis of no effect for lower teeth.

In parallel, using $\psi_{\text{in}}(\cdot)$ rather than $\psi_{\text{hu}}(\cdot)$, the intervals of point estimates and the 95% confidence intervals are further from zero. At $\Gamma = 1.5$, as above, but using $\psi_{\text{in}}(\cdot)$ rather than $\psi_{\text{hu}}(\cdot)$, the interval of point estimates of a shift for lower teeth is $[4.85, 9.73]$ and the 95% confidence interval is $[3.28, 11.71]$, whereas for upper teeth the point estimates are $[2.60, 6.73]$ and the confidence interval is $[1.41, 8.76]$.

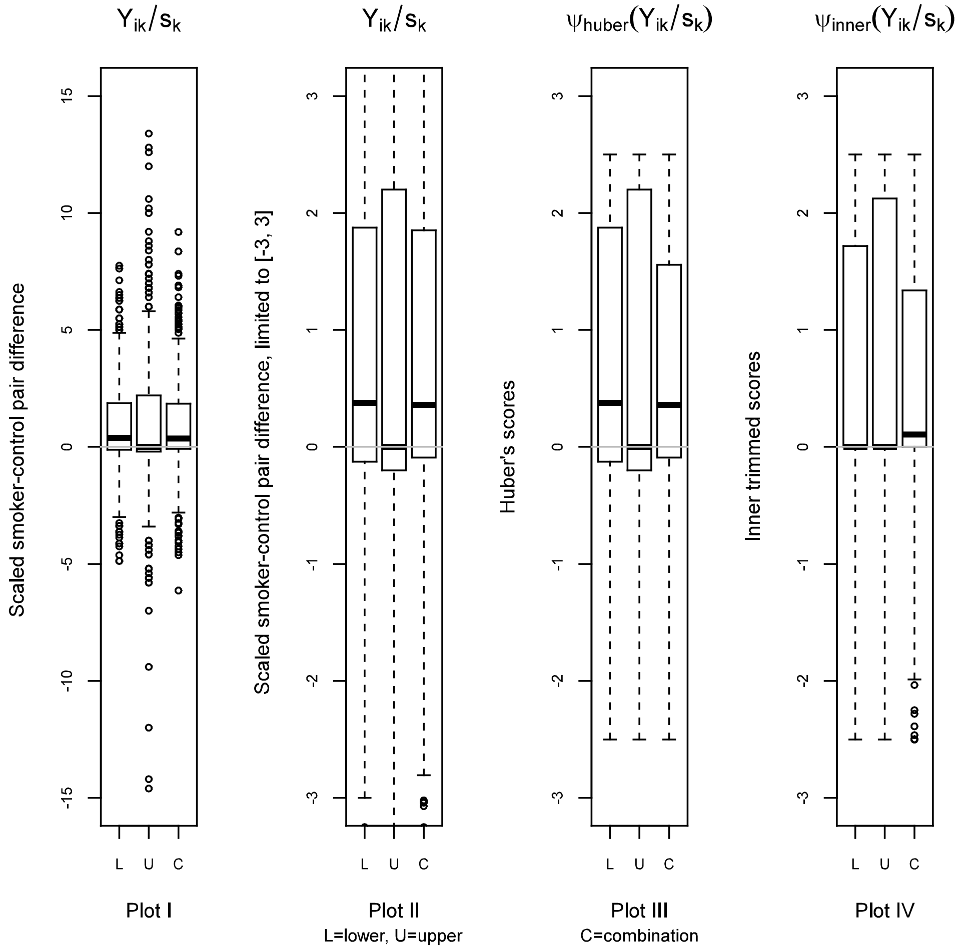

FIG. 3. *Smoker-minus-control matched pair differences for lower teeth* (*L*), *upper teeth* (*U*), *and the linear combination* $\lambda = (0.714, 0.286)^{\mathrm{T}}$ *that gives more than twice the weight to lower teeth. All values are scaled*, $s_1 = 8$, $s_2 = 5$. *Plot* I *has an unrestricted range*, *but the range of Plots* II, III, *and* IV *are restricted to* $[-3, 3]$. *Plot* III *uses Huber's* ψ_{hu} *and Plot* IV *uses inner trimming*, ψ_{in}. *With Huber's* ψ_{hu} *in Plot* III, *values above* 2.5 *become* 2.5, *and values below* -2.5 *become* -2.5. *With inner trimming* ψ_{in} *in Plot* IV, *additionally*, *values between* -0.5 *and* 0.5 *become* 0. *In each plot*, *there is a horizontal line at* 0.

Figure 3 compares $\psi_{\mathrm{hu}}(\cdot)$ and $\psi_{\mathrm{in}}(\cdot)$ for the pair differences for lower teeth, Y_{i1}, upper teeth, Y_{i2}, and the linear combination with weights $\boldsymbol{\lambda} = (0.714, 0.286)^T$. Plots II, III and IV of Figure 3 restrict the y-axis to the interval $[-3, 3]$ so that the extremes do not obscure the center of the plot. In the M-statistics, the weights are applied after scaling the responses, so Figure 3 plots Y_{ik}/s_k, $\psi_{\mathrm{hu}}(Y_{ik}/s_k)$ and $\psi_{\mathrm{in}}(Y_{ik}/s_k)$ and their linear combinations weighted by $\boldsymbol{\lambda} = (0.714, 0.286)^T$. Huber's $\psi_{\mathrm{hu}}(\cdot)$ in Plot III replaces values beyond ±2.5 by ±2.5. Inner trimming in

Plot IV does this and additionally replaces values in $[-0.5, 0.5]$ by zero. The combined scores $0.714\psi(Y_{i1}/s_1) + 0.286\psi(Y_{i2}/s_2)$ in Plots III and IV have shorter interquartile distances than the separate components. Figure 3 provides a visual display of the numerical finding above that the deviate in (5.1) was least sensitive for an unequally weighted combination of lower and upper teeth using $\psi_{\text{in}}(\cdot)$.

5.3. *Analyses restricted to heavier smokers.* Among the 441 daily smokers, the median number of cigarettes smoked per day was 10, and $206/441 = 47\%$ smoked more than 10 per day. Results are insensitive to somewhat larger unmeasured biases if the analysis is restricted to the 206 pairs in Figures 2 and 3 in which the smoker smoked more than 10 cigarettes per day. Using $\psi_{\text{in}}(\cdot)$, the corrected P-value from Proposition 1 is 0.0494 at $\Gamma = 2.55$ for these 206 pairs, although the optimal weights now attach weight 1 to lower teeth and 0 to upper teeth, $\boldsymbol{\lambda} = (1.0, 0.0)^T$. In contrast, using all teeth, $\boldsymbol{\lambda} = (0.5, 0.5)$, the adjusted P-value from Proposition 1 is 0.308 at $\Gamma = 2.55$. For general results about design sensitivities when low-dose pairs are eliminated, see Rosenbaum (2010), Section 17.3.

5.4. *Sensitivity analyses using scheffé projections.* It would take a moderately large bias from unmeasured covariates to explain away the ostensible effects of smoking on periodontal disease. There is evidence of an effect for both lower and upper teeth, but the ostensible effect on lower teeth is larger than on upper teeth, and the effect on lower teeth is insensitive to larger biases. In the example, the contrast $\boldsymbol{\lambda}$ that produced the least sensitive finding varied from one analysis to another. When looking at all 441 daily smokers, it was best to combine lower and upper teeth, emphasizing lower teeth. When looking at the 206 smokers who smoked more than 10 cigarettes per day, it was best to focus exclusively on lower teeth, ignoring upper teeth. In Tables 2 and 3 and in the example, it was often worthwhile to pay a price for multiple testing to gain the freedom to look at every possible contrast $\boldsymbol{\lambda}$, rather than to try to guess the best $\boldsymbol{\lambda}$ from a priori considerations. The design sensitivities in Table 2 indicate that the correct choice of $\boldsymbol{\lambda}$ becomes ever more important, and the multiplicity correction ever less important, as the sample size I increases.

6. Discussion: Multivariate outcomes viewed as infinitely many univariate outcomes.

When a large observational study has a K-dimensional outcome, a test that considers all linear combinations of the K-dimensional outcome can exhibit greater insensitivity to unmeasured biases than a test that considers the K outcomes one at a time with a correction for multiple testing. More precisely, consideration of all linear combinations of a K-dimensional outcome cannot reduce but can increase the design sensitivity, the limiting sensitivity to unmeasured bias as the sample size increases, $I \to \infty$. Gains can occur when several outcomes are each affected by the treatment or when one outcome is strongly affected and another correlated outcome is entirely unaffected. In practice, gains in the power of a sensitivity analysis combined with gains in understanding are most likely to occur when: (i) the sample size I is reasonably large and the dimension K is not large, (ii) the K-outcomes are components of a whole, such as upper and lower teeth in one mouth, a few subscores of a test score or attitude scale, or measures of related changes in behavior.

In the periodontal data, smoking appeared to cause periodontal disease, but the linear combination $\boldsymbol{\lambda}$ of results for upper and lower teeth that yielded the greatest insensitivity to unmeasured biases varied from one analysis to another. Analyses that used all smokers found greatest insensitivity with a $\boldsymbol{\lambda}$ that used both lower and upper teeth, but with lower teeth receiving more than twice the weight of upper teeth. Analyses that focused on heavier smokers were most insensitive if upper teeth were ignored entirely. It would be difficult to anticipate these patterns before looking at the data. Use of Scheffé projections permits the investigator to search for a particularly insensitive combination $\boldsymbol{\lambda}$ while controlling the probability of falsely rejecting a true hypothesis.

REFERENCES

BILLINGSLEY, P. (1979). *Probability and Measure*. Wiley, New York. MR0534323

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. *J. Natl. Cancer Inst.* **22** 173–203.

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

FOGARTY, C. B. and SMALL, D. S. (2016). Sensitivity analysis for multiple comparisons in matched observational studies through quadratically constrained linear programming. *J. Amer. Statist. Assoc.* To appear, DOI:10.1080/01621459.2015.1120675.

GASTWIRTH, J. L., KRIEGER, A. M. and ROSENBAUM, P. R. (2000). Asymptotic separability in sensitivity analysis. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 545–555. MR1772414

GENZ, A. and BRETZ, F. (2009). *Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics* **195**. Springer, Dordrecht. MR2840595

HANSEN, B. B. (2007). Optmatch (R package optmatch). *R News* **7** 18–24.

HELLER, R., ROSENBAUM, P. R. and SMALL, D. S. (2009). Split samples and design sensitivity in observational studies. *J. Amer. Statist. Assoc.* **104** 1090–1101. MR2750238

HODGES, J. L. JR. and LEHMANN, E. L. (1962). Rank methods for combination of independent experiments in analysis of variance. *Ann. Math. Stat.* **33** 482–497. MR0156426

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

HOSMAN, C. A., HANSEN, B. B. and HOLLAND, P. W. (2010). The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Stat.* **4** 849–870. MR2758424

HSU, J. Y. and SMALL, D. S. (2013). Calibrating sensitivity analyses to observed covariates in observational studies. *Biometrics* **69** 803–811. MR3146776

HUBER, P. J. (1981). *Robust Statistics*. Wiley, New York. MR0606374

KARLIN, S. (1992). *Mathematical Methods and Theory in Games, Programming, and Economics*. Dover Publications, New York. Vol. I: Matrix games, programming, and mathematical economics, Vol. II: The theory of infinite games, Reprint of the 1959 original. MR1160778

LEHMACHER, W., WASSMER, G. and REITMEIR, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics* **47** 511–521.

LEHMANN, E. L. (1975). *Nonparametrics*. Holden Day, San Francisco.

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

LIU, W., KURAMOTO, S. J. and STUART, E. A. (2013). An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prev. Sci.* **14** 570–580.

MANSKI, C. F. (1990). Nonparametric bounds on treatment effects. *Am. Econ. Rev.* **80** 319–323.

MANSKI, C. F. and NAGIN, D. S. (1990). Bounding disagreements about treatment effects: A case study of sentencing and recidivism. *Sociol. Method.* **28** 99–137.

MARCUS, R., PERITZ, E. and GABRIEL, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056

MARITZ, J. S. (1979). A note on exact robust confidence intervals for location. *Biometrika* **66** 163–166. MR0529161

MCCANDLESS, L. C., GUSTAFSON, P. and LEVY, A. (2007). Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Stat. Med.* **26** 2331–2347. MR2368419

MCKILLIP, J. (1992). Research without control groups: A control construct design. In *Methodological Issues in Applied Social Psychology* (F. B. Bryant et al., eds.) 159–175. Plenum Press, New York.

NEYMAN, J. (1923, 1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9 [Originally published in Polish in *Ann. Agric. Sci.* **10** (1923), 1–51]. Reprinted in 1990 in *Statist. Sci.* **5** 463–464.

PITMAN, E. J. G. (1937). Significance tests that may be applied to samples from any population, I. *Supp. J. Roy. Satist. Soc.* **4** 119–130.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York. MR0346957

ROSENBAUM, P. R. (1995). Quantiles in nonrandom samples and observational studies. *J. Amer. Statist. Assoc.* **90** 1424–1431. MR1379486

ROSENBAUM, P. R. (2004). Design sensitivity in observational studies. *Biometrika* **91** 153–164. MR2050466

ROSENBAUM, P. R. (2007). Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. *Biometrics* **63** 456–464. MR2370804

ROSENBAUM, P. R. (2010). *Design of Observational Studies*. Springer, New York. MR2561612

ROSENBAUM, P. R. (2013). Impact of multiple matched controls on design sensitivity in observational studies. *Biometrics* **69** 118–127. MR3058058

ROSENBAUM, P. R. (2014). Weighted M-statistics with superior design sensitivity in matched observational studies with multiple controls. *J. Amer. Statist. Assoc.* **109** 1145–1158. MR3265687

ROSENBAUM, P. R. (2015). Bahadur efficiency of sensitivity analyses in observational studies. *J. Amer. Statist. Assoc.* **110** 205–217. MR3338497

ROSENBAUM, P. R. and RUBIN, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Satist. Soc. B* **45** 212–218.

ROSENBAUM, P. R. and SILBER, J. H. (2009a). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405. MR2750570

ROSENBAUM, P. R. and SILBER, J. H. (2009b). Sensitivity analysis for equivalence and difference in an observational study of neonatal intensive care units. *J. Amer. Statist. Assoc.* **104** 501–511. MR2751434

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Ed. Psych.* **66** 688–701.

SCHEFFÉ, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika* **40** 87–104. MR0057504

SHEPHERD, B. E., GILBERT, P. B., JEMIAI, Y. and ROTNITZKY, A. (2006). Sensitivity analyses comparing outcomes only existing in a subset selected post-randomization, conditional on covariates, with application to HIV vaccine trials. *Biometrics* **62** 332–342. MR2236845

STUART, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statist. Sci.* **25** 1–21. MR2741812

TOMAR, S. L. and ASMA, S. (2000). Smoking-attributable periodontitis in the United States: Findings from NHANES III. *J. Periodont.* **71** 743–751.

VAN ELTEREN, PH. (1960). On the combination of independent two sample test of Wilcoxon. *Bull. Inst. Internat. Statist.* **37** 351–361. MR0119313

WEI, L., BARKER, L. and EKE, P. (2013). Array applications in determining periodontal disease measurement. SouthEast SAS User's Group. (SESUG2013) Paper CC-15, analytics.ncsu.edu/sesug/2013/CC-15.pdf.

WEISS, N. (2002). Can the 'specificity' of an association be rehabilitated as a basis for supporting a causal hypothesis? *Epidemiology* **13** 6–8.

WELCH, B. L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika* **29** 21–52.

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

ZUBIZARRETA, J. R. (2012). Using mixed integer programming for matching in an observational study of kidney failure after surgery. *J. Amer. Statist. Assoc.* **107** 1360–1371. MR3036400

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6340
USA
E-MAIL: rosenbaum@wharton.upenn.edu