2012

An Exact Adaptive Test With Superior Design Sensitivity in an Observational Study of Treatments for Ovarian Cancer

Paul R. Rosenbaum
University of Pennsylvania

Recommended Citation

Rosenbaum, P. R. (2012). An Exact Adaptive Test With Superior Design Sensitivity in an Observational Study of Treatments for Ovarian Cancer. *Annals of Applied Statistics, 6* (1), 83-105. http://dx.doi.org/10.1214/11-AOAS508

An Exact Adaptive Test With Superior Design Sensitivity in an Observational Study of Treatments for Ovarian Cancer

Abstract

A sensitivity analysis in an observational study determines the magnitude of bias from nonrandom treatment assignment that would need to be present to alter the qualitative conclusions of a naïve analysis that presumes all biases were removed by matching or by other analytic adjustments. The power of a sensitivity analysis and the design sensitivity anticipate the outcome of a sensitivity analysis under an assumed model for the generation of the data. It is known that the power of a sensitivity analysis is affected by the choice of test statistic, and, in particular, that a statistic with good Pitman efficiency in a randomized experiment, such as Wilcoxon's signed rank statistic, may have low power in a sensitivity analysis and low design sensitivity when compared to other statistics. For instance, for an additive treatment effect and errors that are Normal or logistic or *t*-distributed with 3 degrees of freedom, Brown's combined quantile average test has Pitman efficiency close to that of Wilcoxon's test but has higher power in a sensitivity analysis, while a version of Noether's test has poor Pitman efficiency in a randomized experiment but much higher design sensitivity so it is vastly more powerful than Wilcoxon's statistic in a sensitivity analysis if the sample size is sufficiently large. A new exact distribution-free test is proposed that rejects if either Brown's test or Noether's test rejects after adjusting the two critical values so the overall level of the combined test remains at α, conventionally $\alpha = 0.05$. In every sampling situation, the design sensitivity of the adaptive test equals the larger of the two design sensitivities of the component tests. The adaptive test exhibits good power in sensitivity analyses asymptotically and in simulations. In one sampling situation—Normal errors and an additive effect that is three-quarters of the standard deviation with 500 matched pairs—the power of Wilcoxon's test in a sensitivity analysis was 2% and the power of the adaptive test was 87%. A study of treatments for ovarian cancer in the Medicare population is discussed in detail.

Keywords

Brown's test, combined quantile averages, design sensitivity, Noether's Test, observational study, randomization inference, sensitivity analysis, Wilcoxon's signed rank test

Disciplines

Applied Statistics

The Annals of Applied Statistics
2012, Vol. 6, No. 1, 83–105
DOI: 10.1214/11-AOAS508

AN EXACT ADAPTIVE TEST WITH SUPERIOR DESIGN SENSITIVITY IN AN OBSERVATIONAL STUDY OF TREATMENTS FOR OVARIAN CANCER

BY PAUL R. ROSENBAUM[1]

University of Pennsylvania

A sensitivity analysis in an observational study determines the magnitude of bias from nonrandom treatment assignment that would need to be present to alter the qualitative conclusions of a naïve analysis that presumes all biases were removed by matching or by other analytic adjustments. The power of a sensitivity analysis and the design sensitivity anticipate the outcome of a sensitivity analysis under an assumed model for the generation of the data. It is known that the power of a sensitivity analysis is affected by the choice of test statistic, and, in particular, that a statistic with good Pitman efficiency in a randomized experiment, such as Wilcoxon's signed rank statistic, may have low power in a sensitivity analysis and low design sensitivity when compared to other statistics. For instance, for an additive treatment effect and errors that are Normal or logistic or t-distributed with 3 degrees of freedom, Brown's combined quantile average test has Pitman efficiency close to that of Wilcoxon's test but has higher power in a sensitivity analysis, while a version of Noether's test has poor Pitman efficiency in a randomized experiment but much higher design sensitivity so it is vastly more powerful than Wilcoxon's statistic in a sensitivity analysis if the sample size is sufficiently large. A new exact distribution-free test is proposed that rejects if either Brown's test or Noether's test rejects after adjusting the two critical values so the overall level of the combined test remains at α, conventionally $\alpha = 0.05$. In every sampling situation, the design sensitivity of the adaptive test equals the larger of the two design sensitivities of the component tests. The adaptive test exhibits good power in sensitivity analyses asymptotically and in simulations. In one sampling situation—Normal errors and an additive effect that is three-quarters of the standard deviation with 500 matched pairs—the power of Wilcoxon's test in a sensitivity analysis was 2% and the power of the adaptive test was 87%. A study of treatments for ovarian cancer in the Medicare population is discussed in detail.

1. Introduction: Motivation; example; outline.

1.1. *Are large observational studies less susceptible to unmeasured biases*? There is certainly a sense in which large observational studies are more—not

Received May 2011; revised August 2011.

[1]Supported by a grant from the NSF.

Key words and phrases. Brown's test, combined quantile averages, design sensitivity, Noether's test, observational study, randomization inference, sensitivity analysis, Wilcoxon's signed rank test.

less—susceptible to unmeasured biases than smaller studies. Biases due to nonrandom treatment assignment generally do not become smaller as the sample size increases. These biases are due to the failure to control some unmeasured covariate that would have been balanced by random assignment of treatments. If a large observational study is analyzed naïvely under the assumption that adjustments for measured covariates have, in effect, transformed the study into a randomized experiment, then as the sample size increases even very small biases due to unmeasured covariates can seriously distort the level of significance tests and the coverage of confidence intervals; see Cochran (1965), Section 3.1.

Suppose, however, that the analysis takes explicit account of uncertainty about unmeasured biases by performing a sensitivity analysis. Is a large sample size of any assistance in this case? It is known that the degree of sensitivity to unmeasured biases is affected by many aspects of the design and analysis of an observational study [Rosenbaum (2004, 2010b)], but the relevant decisions about design and analysis are often difficult to make without guidance from empirical data. Heller, Rosenbaum and Small (2009) found that sample splitting—sacrificing a small portion, say, 10%, of the sample to guide design and analysis—could, in favorable circumstances, yield reduced sensitivity to unmeasured biases by guiding the needed decisions. Sample splitting has the advantage, emphasized by Cox (1975), of permitting reflection and judgement in light of data without invalidating the formal properties of statistical procedures. However, some questions, such as the thickness of the tails of distributions, are difficult to settle using a small fraction of the sample, and may require guidance from the complete sample. Here, an adaptive test is proposed that chooses between two tests with different properties, and in one sense achieves the performance of the better test in large samples; see Proposition 1 in Section 4.3. Although motivated by large sample calculations, the adaptive procedure performs well in simulations in samples as small as 100 matched pairs.

1.2. *Example*: *Is more chemotherapy for ovarian cancer more effective*? Following surgery to remove a visible tumor, the typical reason that one cancer patient receives more chemotherapy than another is that their cancers differ in localization or recurrence. A straightforward comparison of patients receiving more or less chemotherapy is likely to be biased by comparing sicker patients to healthier ones. Is there a better comparison? Ovarian cancer is unusual in this regard, because there is a source of variation in the intensity of chemotherapy that is not a reaction to the patient and her illness. Chemotherapy for ovarian cancer may be provided by either a medical oncologist who treats cancers of all kinds or by a gynecological oncologist who treats cancers of the ovary, uterus and cervix. Medical oncologists (MOs) and gynecological oncologists (GOs) differ in both training and practice. In particular, GOs are gynecologists, and hence surgeons, perhaps the best surgeons for gynecological cancers, and they often perform surgery for ovarian cancer, whereas MOs are almost invariably not surgeons and administer chemotherapy after someone else, perhaps a general surgeon, a gynecologist or GO, has performed

surgery. Typically, an MO had a residency in internal medicine followed by a 3-year fellowship in oncology emphasizing the use of chemotherapy, whereas a GO had a residency in obstetrics and gynecology followed by a fellowship in gynecologic oncology with attention paid to surgical treatment of ovarian cancer. Silber et al. (2007) hypothesized correctly that MOs would use chemotherapy more intensively than GOs, and they used this difference in intensity to ask whether more chemotherapy is of benefit to the patient.

Using data from Medicare and the Surveillance, Epidemiology and End Results (SEER) program of the U.S. National Cancer Institute, Silber et al. (2007) looked at patients with ovarian cancer between 1991 and 2001 who received chemotherapy after appropriate surgery; see their paper for details of the patient population. They matched all $I = 344$ such ovarian cancer patients treated by a gynecologic oncologist to 344 ovarian cancer patients treated by a medical oncologist. Using the matching algorithm of Rosenbaum, Ross and Silber (2007), the matching controlled for 36 covariates, including clinical stage, tumor grade, surgeon type, comorbid conditions such as diabetes and congestive heart failure, age, race and year of diagnosis [Silber et al. (2007), Tables 2 and 3]. Importantly, the duration of follow-up was virtually identical in the two groups. On average, during the five years after diagnosis, the patients of medical oncologists received about four more weeks of chemotherapy, with MO patients receiving on average 16.5 weeks of chemotherapy and GO patients receiving 12.1 weeks. The upper portion of Figure 1 is a pair of two quantile–quantile plots [Wilk and Gnanadesikan (1968)] of weeks of chemotherapy in the first year or the first five years for the 344 GO patients and the 344 MO patients, momentarily ignoring who is matched to whom. Because the points lie above the line of equality, the distribution of chemotherapy weeks for MO patients appears to be stochastically larger than the distribution for GO patients. Survival was virtually identical with nearly identical Kaplan–Meier survival curves that crossed repeatedly, and a median survival of 2.98 years in the MO group and 3.04 years in the GO group [Silber et al. (2007), Figure 1 and Table 1]. Patients of medical oncologists experienced more weeks with chemotherapy associated side effects or toxicity, such as anemia, neutropenia, thrombocytopenia and drug induced neuropathy, on average over five years, 16.2 weeks for MOs and 8.9 weeks for GOs; see the bottom half of Figure 2. If Wilcoxon's signed rank test is used to compare weeks with toxicity in matched pairs, the P-values are less than 10^{-6} for both year one and the first five years, but of course those P-values take no account of possible biases in this nonrandomized comparison. In brief, greater intensity of chemotherapy was not associated with longer survival, but it was associated with more frequent side effects.

The study generated some discussion, in particular, an editorial, five letters discussing either the study or the editorial, and two rejoinders, one from the authors of the paper and one from the author of the editorial, or 11 pages of published discussion of a 7 page paper. Happily, matching for 36 measured covariates was convincing in the very limited sense such adjustments can be convincing: none of

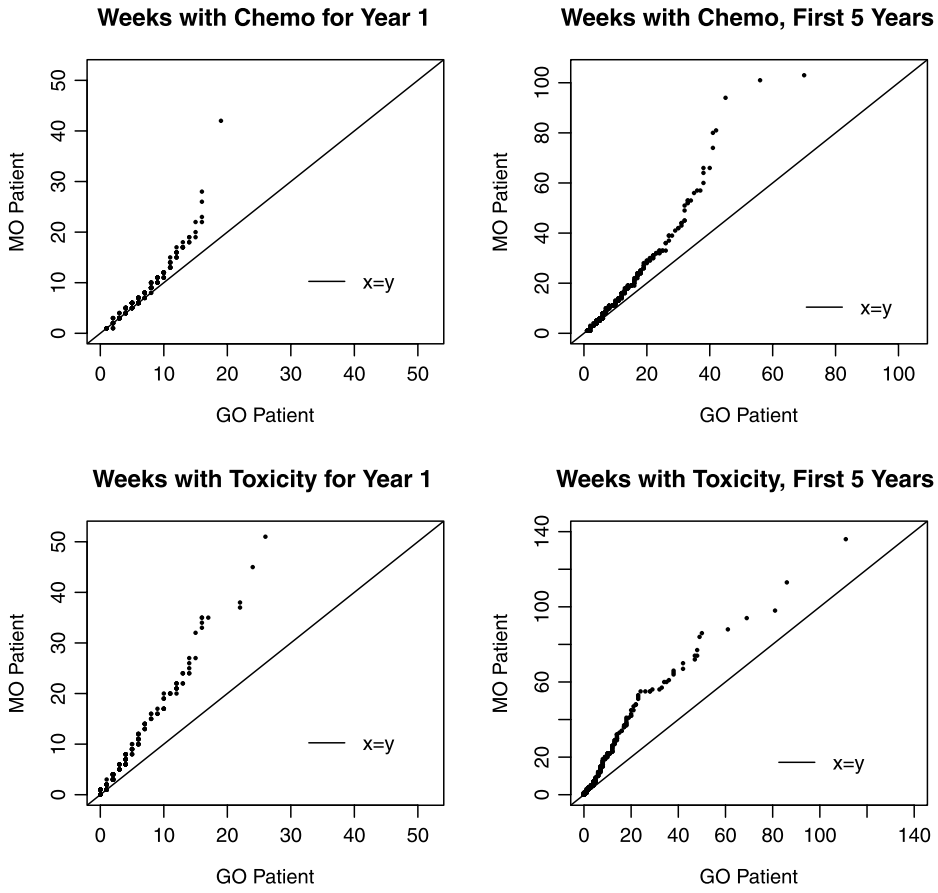

FIG. 1. *Four quantile–quantile plots of weeks with either chemotherapy or toxicity for* $I = 344$ *pairs of an MO and a GO patient. Quantile–quantile plots describe the marginal distributions, ignoring who is paired with whom.*

the discussion expressed continued concern about these 36 measured covariates, which include many of the key covariates for ovarian cancer. Virtually all of the discussion concerned possible unmeasured biases, possible ways the MO and GO groups may differ besides the 36 measured covariates. The editorial by an MO mentioned the magnitude of "residual tumor" not removed by surgery, a covariate not recorded in SEER, and suggested the possibility that GOs were less prone to notice toxicity, whereas the first letter by two GOs characterized these comments as "spinning a tale." The particulars of the discussion had strengths and weaknesses, but, in an abstract sense, a concern about possible unmeasured biases is reasonable in most if not all observational studies, and in that sense the discussion was constructively focused on the central issue. A disappointing feature of the 11 pages of discussion was that it contained little in the way of data, quantitative analysis

or evidence, although there was a little data in one rejoinder. A sensitivity analysis in an observational study is an attempt to return to the data and to quantitative analysis when discussing the possible impact of unmeasured biases.

How large would the departure from random assignment have to be to alter the conclusions? The answer is determined by a sensitivity analysis. The degree of sensitivity to unmeasured biases in this study is noticeably affected by the choice of test statistic; see Section 5. Theoretical considerations suggest that certain statistics, for instance, Wilcoxon's statistic, tend to exaggerate the degree of sensitivity to unmeasured biases, at least for additive treatment effects with symmetric errors [Rosenbaum (2010a)], so perhaps certain methods may be excluded on purely theoretical grounds. On the other hand, many issues affect the degree of sensitivity to bias reported by different test statistics [Rosenbaum (2010b), Part III], and some of these issues are difficult to evaluate prior to looking at the data. Here, an exact, adaptive test is proposed that chooses, after the fact, the less sensitive of two analyses, exactly correcting the level of the test for the use of two analyses. Is adapting the test statistic to the data at hand of value in sensitivity analyses?

1.3. *Outline*: *Review*; *an exact adaptive test*; *design sensitivity*; *power.* Section 2 is a review of existing background material and notation, including randomization inference in experiments in Section 2.1, sensitivity analysis in observational studies in Section 2.2, and the power of a sensitivity analysis and the design sensitivity in Section 2.3; there is little new material in the review in Section 2. With notation and background established, Section 3 discusses why adaptation is important in this context. The new adaptive test is discussed in Section 4, its exact null distribution in Sections 4.1–4.2, its nonnull asymptotic properties in Section 4.3, and its finite sample power obtained by simulation in Section 4.4. In particular, Proposition 1 of Section 4.3 shows that in each sampling situation, the design sensitivity of the adaptive procedure is equal to the maximum of the design sensitivities of the two nonadaptive procedures from which it is built. The simulation suggests that the asymptotic properties begin to take effect in samples of modest size. In Section 5 the methods are applied to the example in Section 1.2 from Silber et al. (2007). The discussion in Section 6 considers related alternative methods in Section 6.1 and returns in Section 6.2 to the question raised in Section 1.1.

2. Notation and review: Randomization; sensitivity analysis; design sensitivity.

2.1. *Inference about treatment effects in a randomized experiment.* There are I matched pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, one treated with $Z_{ij} = 1$, the other control with $Z_{ij} = 0$, so $Z_{i1} + Z_{i2} = 1$ for each i. The subjects were matched for an observed covariate, $\mathbf{x}_{ij}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each i, but they may differ in terms of an unobserved covariate u_{ij}, so possibly $u_{i1} \neq u_{i2}$. Subject ij has two potential responses, namely, r_{Tij} if ij is assigned to treatment with

$Z_{ij} = 1$ and r_{Cij} if ij is assigned to control with $Z_{ij} = 0$, so the response observed from ij is $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$ and the effect of the treatment $r_{Tij} - r_{Cij}$ on subject ij is not observed for any subject; see Neyman (1923), Welch (1937), Rubin (1974), Reiter (2000) and Gadbury (2001). Fisher's (1935) sharp null hypothesis H_0 of no treatment effect asserts $H_0 : r_{Tij} = r_{Cij}$, for $i = 1, \ldots, I$, $j = 1, 2$, whereas the hypothesis H_τ of an additive constant treatment effect τ asserts $H_\tau : r_{Tij} = r_{Cij} + \tau$ for all ij.

Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1, \ldots, I, j = 1, 2\}$ for the potential responses and covariates and write $\mathcal{Z}$ for the event that $(Z_{i1} + Z_{i2} = 1, i = 1, \ldots, I)$. In a randomized paired experiment, one subject in each pair i is picked at random to receive the treatment, so $\Pr(Z_{ij} = 1|\mathcal{F}, \mathcal{Z}) = \frac{1}{2}$ for each ij, with independent assignments in distinct pairs.

Within pair i, the treated-minus-control difference Y_i in observed responses is

$$\begin{aligned} Y_i &= (Z_{i1} - Z_{i2})(R_{i1} - R_{i2}) = Z_{i1}(r_{Ti1} - r_{Ci2}) + Z_{i2}(r_{Ti2} - r_{Ci1}) \\ &= \tau + \varepsilon_i \qquad \text{with } \varepsilon_i = (Z_{i1} - Z_{i2})(r_{Ci1} - r_{Ci2}) \qquad \text{if } H_\tau \text{ is true.} \end{aligned}$$

Given $\mathcal{F}, \mathcal{Z}$, the quantity $r_{Ci1} - r_{Ci2}$ is fixed, and in a randomized experiment $\varepsilon_i = \pm|r_{Ci1} - r_{Ci2}|$ with equal probabilities $\frac{1}{2}$, so if H_τ is true, then Y_i is symmetric about τ.

Ties among the Y_i's are not a problem, but the development is simpler if ties of all kinds are assumed absent. In particular, when testing H_τ, the $|Y_i - \tau|$ are assumed to be untied, and the $Y_i - \tau$ are assumed to not equal zero. Minor adjustments in Section 4.5 eliminate these restrictions.

When testing H_τ, let q_i be the rank of $|Y_i - \tau|$ and let $S_i = 1$ if $Y_i - \tau > 0$ or $S_i = 0$ if $Y_i - \tau \leq 0$; then Wilcoxon's signed rank statistic is $W = \sum_{i=1}^{I} S_i q_i$, where the q_i are a permutation of $1, 2, \ldots, I$. Conditionally given $\mathcal{F}, \mathcal{Z}$, if H_τ is true in a randomized paired experiment, then $Y_i - \tau = \varepsilon_i$ is $\pm|r_{Ci1} - r_{Ci2}|$ with equal probabilities $\frac{1}{2}$, so q_i is fixed and $S_i = 1$ or 0 with equal probabilities $\frac{1}{2}$, and Wilcoxon's statistic has the distribution of the sum of I independent random variables taking the values i or 0 with equal probabilities $\frac{1}{2}$. This null distribution is the basis for testing H_τ, and by inverting the test it yields confidence intervals and Hodges–Lehmann point estimates for an additive treatment effect τ. See Lehmann (1975) for discussion of these standard techniques and for discussion of the good performance of Wilcoxon's statistic when applied in randomized experiments. See Maritz (1979) for a parallel development of randomization inferences using Huber's m-estimates including the permutational t-test.

2.2. *Sensitivity analysis in observational studies.* In the absence of randomization, there is no basis for assuming that $\Pr(Z_{ij} = 1|\mathcal{F}, \mathcal{Z}) = \frac{1}{2}$ and therefore no basis beyond naïveté for assuming that the inferences in Section 2.1 are correct. A sensitivity analysis in an observational study asks how the conclusions in Section 2.1 would change in response to departures from $\Pr(Z_{ij} = 1|\mathcal{F}, \mathcal{Z}) = \frac{1}{2}$

of various magnitudes. One model for sensitivity analysis [Rosenbaum (2002a), Section 4] begins by assuming that in the population before matching treatment assignments are independent with unknown probabilities $\pi_{ij} = \Pr(Z_{ij} = 1|\mathcal{F})$, and two subjects with, say, ij and ij', with the same observed covariates, $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, may differ in their odds of treatment by at most a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1-\pi_{ij'})}{\pi_{ij'}(1-\pi_{ij})} \leq \Gamma \qquad \text{if } \mathbf{x}_{i1} = \mathbf{x}_{i2}; \tag{1}$$

then a distribution of treatment assignments for matched pairs is obtained by conditioning on the event $\mathcal{Z}$. This is easily seen to be equivalent to assuming that

$$\frac{1}{1+\Gamma} \leq \Pr(Z_{i1} = 1|\mathcal{F}, \mathcal{Z}) \leq \frac{\Gamma}{1+\Gamma}, \qquad Z_{i2} = 1 - Z_{i1}, \qquad i = 1, \ldots, I, \tag{2}$$

with independent assignments in distinct pairs; see Rosenbaum (2002a), Section 4. To aid interpretation, the one parameter Γ may be unpacked into two parameters, one Λ controlling the relationship between treatment assignment Z_{ij} and the unobserved covariate u_{ij}, the other Δ controlling the relation between response r_{Cij} and u_{ij}, yielding the same one-dimensional analysis in terms of Γ but for all (Λ, Δ) that solve $\Gamma = (\Lambda\Delta + 1)/(\Lambda + \Delta)$ [Rosenbaum and Silber (2009)]; for instance, $\Gamma = 1.25$ corresponds with an unobserved covariate that simultaneously doubles the odds of treatment, $Z_{i1} - Z_{i2} = 1$, and doubles the odds of a positive response difference under control, $r_{Ci1} - r_{Ci2} > 0$, as $1.25 = (2 \times 2 + 1)/(2 + 2)$. In this formulation, the parameter Δ is defined using Wolfe's (1974) semiparametric family of asymmetric deformations of a symmetric distribution to place a bound on the distribution of $r_{Ci1} - r_{Ci2}$; see Rosenbaum and Silber (2009) for specifics. Either (1) or (2) says that treatment assignment probabilities are unknown but to a bounded degree determined by Γ. For each fixed $\Gamma \geq 1$, (2) yields an interval of possible values of an inference quantity, such as a P-value or point estimate or the endpoint of a confidence interval, and a sensitivity analysis consists in computing that interval for several values of Γ, thereby indicating the magnitude of departure from randomization that would need to be present to alter the conclusions of the analysis in Section 2.1. For instance, for $\Gamma = 1$ the interval of one-sided P-values from Wilcoxon's test is a single point, namely, the P-value from the randomization test in Section 2.1, but as $\Gamma \to \infty$ the interval tends to $[0, 1]$—that is, association does not logically imply causation. The practical question is: how large must Γ be before the interval of P-values is inconclusive, say, including values both above and below a conventional level such as 0.05?

Various methods of sensitivity analysis in observational studies are discussed by Cornfield et al. (1959), Copas and Eguchi (2001), Diprete and Gangl (2004), Egleston, Scharfstein and MacKenzie (2009), Frangakis and Rubin (1999), Gastwirth (1992), Gilbert, Bosch and Hudgens (2003), Hosman, Hansen and Holland (2010), Imbens (2003), Lin, Psaty and Kronmal (1998), Marcus (1997), McCandless, Gustafson and Levy (2007), Rosenbaum and Rubin (1983), Small

(2007), Wang and Krieger (2006), Yanagawa (1984), Yu and Gastwirth (2005), among others.

The discussion has emphasized adjustments for observed covariates by matching, as opposed to, say, covariance adjustment. In simulations, Rubin (1979) found that model-based adjustments without matching are not robust to model misspecification, sometimes increasing rather than reducing bias from measured covariates, but he found that model-based adjustments of matched pair differences are robust to model misspecification. The methods described in the current paper may be applied to residuals of covariance adjustment of matched pair differences using the device in Rosenbaum (2002b), Section 5. Also, the sensitivity model (1) is applicable to a wide variety of situations, including binary outcomes and censored survival times [Rosenbaum (2002a), Section 4].

2.3. *Design sensitivity in observational studies*. If an observational study were free of unmeasured bias, then we could not determine this from the observable data, and the best we could hope to say is that the conclusions are insensitive to small and moderate biases. The power of a sensitivity analysis is the probability that we will be able to say this [Rosenbaum (2004)]. The power of a randomization test anticipates the outcome of such a test under an assumed model for data generation in a randomized trial. In parallel, the power of a sensitivity analysis with a specific Γ anticipates the outcome of a sensitivity analysis when performed on data from an assumed model for data generation. In the *favorable situation*, the data reflect a treatment effect and no bias from unmeasured covariates, and it is in this situation that we hope to report insensitivity to unmeasured bias. For instance, we might ask the following: if the I matched pair differences were produced by an additive constant treatment effect τ with no bias and Normal errors, $Y_i = \tau + \varepsilon_i$ with $\varepsilon_i \sim_{\text{i.i.d.}} N(0, \sigma^2)$, then, under this model, what is the probability that the entire interval of P-values testing H_0 is below 0.05 when computed with, say, $\Gamma = 2$? For Wilcoxon's test with $I = 100$ and $\tau/\sigma = 1/2$, the entire interval of P-values computed with $\Gamma = 2$ is less than 0.05 with probability 0.54, so there is a reasonable chance that an effect of this magnitude will be judged insensitive to a moderately large bias of $\Gamma = 2$. In contrast, the new adaptive test proposed in the current paper has power of 0.68 in this same situation, a substantial improvement.

As $I \to \infty$, there is a value $\widetilde{\Gamma}$ called the design sensitivity [Rosenbaum (2004)] such that the power of a sensitivity analysis tends to 1 if the analysis is performed with $\Gamma < \widetilde{\Gamma}$ and tends to zero if the analysis is performed with $\Gamma > \widetilde{\Gamma}$. That is, the power, viewed as a function of Γ is tending to a step function with a single step down from 1 to 0 at $\Gamma = \widetilde{\Gamma}$; see Rosenbaum (2010b), Figure 14.3. In the limit, as the sample size increases, data generated by a certain model without bias will be insensitive to biases smaller than $\widetilde{\Gamma}$ and sensitive to biases larger than $\widetilde{\Gamma}$. For instance, if $Y_i = \tau + \varepsilon_i$ with $\varepsilon_i \sim_{\text{i.i.d.}} N(0, \sigma^2)$ and $\tau/\sigma = 1/2$, the design sensitivity for Wilcoxon's statistic is $\widetilde{\Gamma} = 3.17$, so for sufficiently large I it is

virtually certain that Wilcoxon's statistic will report insensitivity to a bias of Γ if $\Gamma < 3.17$ and virtually certain it will report sensitivity to a bias of Γ if $\Gamma > 3.17$. In contrast, in this same sampling situation, the new adaptive test proposed in the current paper has design sensitivity $\widetilde{\Gamma} = 4.97$, again a substantial improvement. In particular, in this sampling situation as $I \to \infty$, the power of a sensitivity analysis performed at $\Gamma = 4$ is tending to zero for Wilcoxon's test and to one for the new adaptive test.

Design sensitivity has been described in terms of the power of tests, but parallel issues arise in conducting a sensitivity analysis for a confidence interval or a point estimate. In a randomized experiment, a test such as Wilcoxon's test may be inverted to yield a confidence interval or a Hodges–Lehmann point estimate of an additive treatment effect τ, and a more powerful test yields a typically shorter confidence interval and more accurate point estimate; see Hodges and Lehmann (1963) or Lehmann (1975), Section 4. In parallel, in an observational study, a sensitivity analysis for a confidence interval or a point estimate is obtained by inverting a test, so the 95% confidence interval for a given Γ excludes τ_0 if the sensitivity analysis for the test rejects $H_0 : \tau = \tau_0$; see Rosenbaum (1993, 2002a), Section 4.3. For a given $\Gamma > 1$, one obtains an interval of possible point estimates and a set of possible confidence intervals. As $I \to \infty$ with $\Gamma > 1$ fixed, both the interval of possible point estimates and the union of possible confidence intervals converges to a real interval $[\tau_L, \tau_H]$ of the treatment effects τ that are compatible with a bias of Γ, and an increase in design sensitivity will shorten that interval; see Rosenbaum [(2005), Proposition 1] for one such result. As in experiments, even if one is interested in a confidence interval or point estimate, not a hypothesis test, one should obtain that interval or estimate by inverting a more powerful test.

3. Why is adaptation important? Traditionally, adaptive methods have selected the best of several statistical procedures using the data at hand and they have focused on improving efficiency in randomized experiments in the absence of bias; see, for instance, Hogg (1974), Policello and Hettmansperger (1976) and Jones (1979). As discussed in Rosenbaum (2010a, 2011), Pitman efficiency and design sensitivity both affect the power of a sensitivity analysis in an observational study, but they can work at cross-purposes. Pitman efficiency aims at power to detect small effects in randomized experiments where bias is not an issue. In an observational study, Pitman efficiency predicts the outcome of a sensitivity analysis for $\Gamma = 1$ in the favorable situation; that is, it predicts the outcome of a randomization test applied in an observational study when bias is eliminated by adjustments such as matching. Small effects, however, are invariably sensitive to small unobserved biases, which are absent in an idealized randomized experiment but can never be excluded from consideration in an observational study. Procedures with superior design sensitivity in observational studies look for stable evidence of moderately large effects, in effect ignoring pairs i with small $|Y_i|$.

There are procedures with good Pitman efficiency and better design sensitivity than Wilcoxon's statistic, and other statistics with poor Pitman efficiency and vastly better design sensitivity than Wilcoxon's statistic. For instance, in testing H_0, Brown (1981) proposed a statistic which ignores the $\frac{1}{3}$ of pairs with the smallest $|Y_i|$ or q_i, gives weight 1 to the signs of the $\frac{1}{3}$ of pairs with the middle values of $|Y_i|$ or q_i, and gives weight 2 to the remaining $\frac{1}{3}$ of pairs with the largest values of $|Y_i|$ or q_i. Brown (1981) shows his statistic is highly robust and almost as efficient as Wilcoxon's statistic in a randomized experiment, whereas in Rosenbaum (2010a) it is seen that Brown's statistic has higher design sensitivity in a range of sampling situations; in combination, these two facts produce improved power in a sensitivity analysis. Noether (1973) proposed a simpler class of statistics that simply counts the number of positive Y_i among pairs with large $|Y_i|$ or q_i. Markowski and Hettmansperger (1982) studied the Pitman efficiency of many statistics similar to those of Brown and Noether by varying the number of pairs that are given various weights; see also the group rank statistics of Gastwirth (1966) and Groeneveld (1972). In the version used in the current paper—but not in Noether's paper—Noether's statistic counts the number of positive Y_i among the $\frac{1}{3}$ of pairs with largest $|Y_i|$ or q_i. Having mentioned once that Noether did not promote this specific version of his statistic, I will not mention this again, and will refer to the statistic as Noether's statistic. Brown's statistic has Pitman efficiency of 0.95 relative to the Wilcoxon statistic for an additive effect with Normal errors, but Noether's statistic has Pitman efficiency of only 0.78, so one would not use this version of Noether's statistic for Normal data from a randomized experiment. Table 1 gives Pitman efficiencies in a paired randomized experiment. In contrast, in a sensitivity analysis in an observational study, if $Y_i = \tau + \varepsilon_i$ with $\varepsilon_i \sim_{\text{i.i.d.}} N(0, \sigma^2)$ and $\tau/\sigma = 1/2$, the design sensitivity for Wilcoxon's statistic is $\widetilde{\Gamma} = 3.17$, for Brown's statistic is $\widetilde{\Gamma} = 3.60$ and for Noether's statistic is $\widetilde{\Gamma} = 4.97$, so for sufficiently large I Noether's statistic will be the best performer in a sensitivity analysis.

TABLE 1

Pitman asymptotic relative efficiency versus the Wilcoxon statistic for a shift alternative in a paired randomized experiment with errors from a Normal distribution, a logistic distribution or a t-distribution with 3 degrees of freedom. In the version used here, Noether's statistic is the number of positive differences among the 1/3 of pairs with the largest absolute differences

	Normal	**Logistic**	***t* 3 df**
Sign	0.67	0.75	0.85
Noether	0.78	0.69	0.59
Brown	0.95	0.94	0.93
Wilcoxon	1.00	1.00	1.00

The adaptive test uses both Brown's statistic and Noether's statistic. Adjusting the critical values to control the level of the test, the adaptive test rejects if either Brown's statistic or Noether's statistic supports rejection. In every sampling situation, the adaptive test has the larger of the two design sensitivities for Brown's and Noether's statistics. The important issue, however, is the power of a sensitivity analysis for finite I. Because asymptotic claims for some adaptive procedures are not readily seen in samples of plausible size, the current paper uses the exact null distribution of the adaptive test and emphasizes finite sample power determined by simulation. The pairing of Brown's statistic and Noether's statistic is a pairing of two strong candidates for which the required exact calculations are feasible.

4. An adaptive test.

4.1. *The exact null distribution in a sensitivity analysis.* Let $0 \leq \lambda_1 < \lambda_2 \leq 1$. Let I_1 be the number of pairs with absolute ranks $q_i \geq (1-\lambda_1)I$ and let B_1 be the number of positive Y_i among these I_1 pairs. Also, let I_2 be the number of ranks with $(1-\lambda_1)I > q_i \geq (1-\lambda_2)I$ and let B_2 be the number of positive Y_i among these I_2 pairs. Noether (1973) proposed B_1 as a test statistic, and Brown (1981) and Markowski and Hettmansperger (1982) proposed $T = 2B_1 + B_2$ as a test statistic; see also Gastwirth (1966).

Let $\overline{\overline{B}}_1$ and $\overline{\overline{B}}_2$ be independent binomials with sample sizes I_1 and I_2 and probabilities of success $\kappa = \Gamma/(1+\Gamma)$, and let $\overline{B}_1$ and $\overline{B}_2$ be independent binomials with sample sizes I_1 and I_2 and probabilities of success $\kappa = 1/(1+\Gamma)$. Also, let $\overline{\overline{T}} = 2\overline{\overline{B}}_1 + \overline{\overline{B}}_2$ and $\overline{T} = 2\overline{B}_1 + \overline{B}_2$. A function $g(\cdot,\cdot)$ is monotone increasing if $g(b_1,b_2) \leq g(b_1',b_2')$ whenever $b_1 \leq b_1'$ and $b_2 \leq b_2'$. Under the sensitivity model (2), if H_0 is true, then it is not difficult to show [Rosenbaum (2002a), Section 4] that for every monotone increasing function $g(\cdot,\cdot)$,

$$\begin{aligned}\Pr\{g(\overline{B}_1,\overline{B}_2) \geq k\} &\leq \Pr\{g(B_1,B_2) \geq k | \mathcal{F},\mathcal{Z}\} \\ &\leq \Pr\{g(\overline{\overline{B}}_1,\overline{\overline{B}}_2) \geq k\} \qquad \text{for every } k,\end{aligned} \tag{3}$$

and the bounds in (3) are sharp in the sense of being attained for some $\Pr(Z_{i1} = 1|\mathcal{F},\mathcal{Z})$ that satisfy (2), so the bounds (3) cannot be improved without additional information that further restricts $\Pr(Z_{i1}=1|\mathcal{F},\mathcal{Z})$. If $\Gamma = 1$, then there is equality throughout (3) and then (3) is the randomization distribution of $g(B_1,B_2)$ under H_0.

Let $g(B_1,B_2) = 1$ if $B_1 \geq k_{B,\Gamma}$ or $2B_1 + B_2 \geq k_{T,\Gamma}$ and $g(B_1,B_2) = 0$ otherwise, for suitable constants $k_{B,\Gamma}$ and $k_{T,\Gamma}$; then $g(\cdot,\cdot)$ is monotone increasing. For a given $\Gamma \geq 1$, the adaptive test rejects H_0 at level α for all π_{ij} satisfying (1) if $B_1 \geq k_{B,\Gamma}$ or $2B_1 + B_2 \geq k_{T,\Gamma}$. The constants $k_{B,\Gamma}$ and $k_{T,\Gamma}$ are determined to satisfy the following conditions:

$$\Pr(\overline{\overline{B}}_1 \geq k_{B,\Gamma} \text{ or } \overline{\overline{T}} \geq k_{T,\Gamma}) \leq \alpha, \tag{4}$$

$$\Pr(\overline{\overline{B}}_1 \geq k_{B,\Gamma} - 1 \text{ or } \overline{\overline{T}} \geq k_{T,\Gamma}) > \alpha \quad \text{and}$$
$$\Pr(\overline{\overline{B}}_1 \geq k_{B,\Gamma} \text{ or } \overline{\overline{T}} \geq k_{T,\Gamma} - 1) > \alpha \tag{5}$$

and

$$|\Pr(\overline{\overline{B}}_1 \geq k_{B,\Gamma}) - \Pr(\overline{\overline{T}} \geq k_{T,\Gamma})| \quad \text{is minimized subject to (4) and (5).} \tag{6}$$

The joint distribution of $(\overline{\overline{B}}_1, \overline{\overline{B}}_2)$ is that of two independent binomials and in R is given by outer(dbinom(0: I_1, I_1, κ), dbinom(0: I_2, I_2, κ), "*"); then finding $k_{B,\Gamma}$ and $k_{T,\Gamma}$ to satisfy (4)–(6) is simply arithmetic.

Although I have never seen this, in principle, there could be two values, $(k_{B,\Gamma}, k_{T,\Gamma})$ and $(k'_{B,\Gamma}, k'_{T,\Gamma}) = (k_{B,\Gamma} - 1, k_{T,\Gamma} + 1)$ that both satisfy (4)–(6). To avoid this ambiguity in the definition of the adaptive procedure, simply use $(k_{B,\Gamma}, k_{T,\Gamma})$ in this extremely unlikely case, thereby preferring to reduce the critical value for Brown's statistic T.

4.2. *Numerical example of the null distribution.* To illustrate the computations in (4)–(6), take $I = 250$ untied pairs, $\Gamma = 4$, $\alpha = 0.05$, and $\lambda_1 = 1/3$, $\lambda_2 = 2/3$; then, $I_1 = 84$, $I_2 = 83$, $\kappa = 4/5$. This yields $k_{B,\Gamma} = 74$ and $k_{T,\Gamma} = 216$ with

$$\Pr(\overline{\overline{B}}_1 \geq 74 \text{ or } \overline{\overline{T}} \geq 216) = 0.0488 \leq \alpha = 0.05, \tag{7}$$

$$\Pr(\overline{\overline{B}}_1 \geq 74) = 0.0370, \qquad \Pr(\overline{\overline{T}} \geq 216) = 0.0320, \tag{8}$$

$$|\Pr(\overline{\overline{B}}_1 \geq 74) - \Pr(\overline{\overline{T}} \geq 216)| = 0.0050. \tag{9}$$

In light of this, for $\Gamma = 4$, the upper bound on the one-sided P-value testing no effect would be less than $\alpha = 0.05$ if either $B_1 \geq 74$ or $T = 2B_1 + B_2 \geq 216$.

Several aspects of the illustration (7)–(9) deserve comment. First, if one were to test using B_1 alone, ignoring B_2, then at $\Gamma = 4$ the upper bound on the one-sided P-value testing no effect would be less than $\alpha = 0.05$ if $B_1 \geq 74$, because $\Pr(\overline{\overline{B}}_1 \geq 74) = 0.0370$ and $\Pr(\overline{\overline{B}}_1 \geq 73) = 0.0691$; that is, in this particular case, owing to the discreteness of the binomial distribution, the adaptive test will reject in every instance in which the test based on B_1 alone rejects and the adaptive test will reject in some other cases as well. Conversely, if one were to test using T alone, then at $\Gamma = 4$ the upper bound on the one-sided P-value testing no effect would be less than $\alpha = 0.05$ if $T \geq 215$ rather than $k_{T,\Gamma} = 216$ in (7) because $\Pr(\overline{\overline{T}} \geq 215) = 0.04288$ and $\Pr(\overline{\overline{T}} \geq 214) = 0.05642$. So, in this one numerical example, the adaptive test rejects in every instance in which the test based on B_1 rejects and also in every instance in which T rejects except $B_1 < 74$ and $T = 215$. Use of the Bonferroni inequality to approximate $\Pr(\overline{\overline{B}}_1 \geq 74 \text{ or } \overline{\overline{T}} \geq 216)$ would err substantially, with $\Pr(\overline{\overline{B}}_1 \geq 74 \text{ or } \overline{\overline{T}} \geq 216) = 0.0488 \leq \Pr(\overline{\overline{B}}_1 \geq 74) + \Pr(\overline{\overline{T}} \geq 216) = 0.0370 + 0.0320 = 0.0690$.

4.3. *Design sensitivity of the adaptive test.* As discussed in Section 2.3, in an observational study, the *favorable situation* means there is a treatment effect and no bias from an unmeasured covariate. In an observational study, we cannot know from the data whether we are in the favorable situation, so the best we can hope to say is that the study's conclusions are insensitive to small and moderate biases. The power of an α-level sensitivity analysis, $0 < \alpha < 1$, performed with a specific $\Gamma \geq 1$, is the probability that the entire interval of possible P-values from the sensitivity analysis is less than or equal to α. For the adaptive test, the power of the sensitivity analysis for fixed Γ is the probability that $B_1 \geq k_{B,\Gamma}$ or $T \geq k_{T,\Gamma}$ when B_1 and T are computed from data that are, in fact, measuring a treatment effect without bias. In principal, one could compute the power conditional upon $\mathcal{F}$, but this would mean that the power would be a function of $\mathcal{F}$, so, in practice, one computes the unconditional power averaging over a simple model for the generation of $\mathcal{F}$. As noted in Section 2.3, the design sensitivity is a number $\widetilde{\Gamma}$ such that the power of an α-level sensitivity analysis tends to 1 as $I \to \infty$ if the sensitivity analysis is performed with $\Gamma < \widetilde{\Gamma}$ and the power tends to zero if the analysis is performed with $\Gamma > \widetilde{\Gamma}$.

In the current paper, the favorable situation refers to treated-minus-control differences Y_i that are drawn independently from a continuous cumulative distribution $F(\cdot)$ that is strictly increasing, $F(y) < F(y')$ if $y < y'$. One of many such favorable situations is $Y_i = \tau + \varepsilon_i$ where the ε_i are independent and identically distributed observations from a continuous, strictly increasing distribution with a density symmetric about zero.

For $y \geq 0$, let $H(y) = F(y) - F(-y)$; then $H(y) = \Pr(|Y_i| \leq y)$, and for $\lambda \in [0, 1)$, the inverse function is well defined with $H^{-1}(\lambda) = y$ if $\lambda = \Pr(|Y_i| \leq y)$. Also define $\zeta(\lambda)$ to be the probability that a Y_i is both positive, $Y_i > 0$, and in the largest λ of the $|Y_i|$, that is, define

$$\zeta(\lambda) = 1 - F\{H^{-1}(1-\lambda)\} = \Pr[(Y_i > 0) \wedge \{|Y_i| > H^{-1}(1-\lambda)\}].$$

Let $\widetilde{\Gamma}_{\mathrm{no}}$, $\widetilde{\Gamma}_{\mathrm{bmh}}$ and $\widetilde{\Gamma}_{\mathrm{ad}}$ be the design sensitivities for, respectively, Noether's statistic B_1, the Brown–Markowski–Hettmansperger statistic T and the adaptive procedure with critical values (4)–(6). That is, B_1 counts the positive Y_i's among the largest λ_1 of the $|Y_i|$, and $T = 2B_1 + B_2$ doubles B_1 and adds the count of the positive Y_i's among the next $\lambda_2 - \lambda_1$ of the $|Y_i|$.

PROPOSITION 1. *If Y_i, $i = 1, \ldots, I$ are independent observations from $F(\cdot)$,*

$$\widetilde{\Gamma}_{\mathrm{no}} = \frac{\zeta(\lambda_1)}{\lambda_1 - \zeta(\lambda_1)}, \tag{10}$$

$$\widetilde{\Gamma}_{\mathrm{bmh}} = \frac{\zeta(\lambda_1) + \zeta(\lambda_2)}{\{\lambda_1 - \zeta(\lambda_1)\} + \{\lambda_2 - \zeta(\lambda_2)\}}, \tag{11}$$

$$\widetilde{\Gamma}_{\mathrm{ad}} = \max(\widetilde{\Gamma}_{\mathrm{no}}, \widetilde{\Gamma}_{\mathrm{bmh}}). \tag{12}$$

PROOF. The proof uses Proposition 2 of Rosenbaum (2010a) which concerns the design sensitivity of a signed rank statistic with general scores; in particular, B_1 and T are two such signed rank statistics. Equations (10) and (11) are obtained by simplifying expression (8) in Proposition 2 of Rosenbaum (2010a), which is a formula for the design sensitivity with general scores. As shown in the proof of that proposition, in a sensitivity analysis performed at a specific value of Γ, the upper bound on the P-value for B_1 converges in probability to zero as $I \to \infty$ if $\Gamma < \widetilde{\Gamma}_{\text{no}}$ and it converges to 1 if $\Gamma > \widetilde{\Gamma}_{\text{no}}$, and, in parallel, the upper bound on the P-value for T converges in probability to zero as $I \to \infty$ if $\Gamma < \widetilde{\Gamma}_{\text{bmh}}$ and it converges to 1 if $\Gamma > \widetilde{\Gamma}_{\text{bmh}}$. As a consequence, the smaller of these two P-values for B_1 and T tends to zero as $I \to \infty$ if $\Gamma < \max(\widetilde{\Gamma}_{\text{no}}, \widetilde{\Gamma}_{\text{bmh}})$ and it tends to 1 if $\Gamma > \max(\widetilde{\Gamma}_{\text{no}}, \widetilde{\Gamma}_{\text{bmh}})$, proving (12). □

Table 2 calculates the design sensitivity of the sign statistic, the Wilcoxon signed rank statistic, Noether's statistic with $\lambda_1 = 1/3$, Brown's statistic with $\lambda_1 = 1/3$, $\lambda_2 = 2/3$, and the adaptive procedure with critical values (4)–(6). In Table 2, $Y_i = \tau + \varepsilon_i$ where $\text{var}(\varepsilon_i) = \sigma^2$, the effect size is specified in units of the standard deviation, τ/σ, and ε_i has a standard Normal distribution, a logistic distribution

TABLE 2
Design sensitivity $\widetilde{\Gamma}$ in the favorable situation with an additive treatment effect, τ and errors ε_i with variance σ^2 that are Normal, logistic or t-distributed with 3 degrees of freedom

Statistic	Normal	Logistic	t 3 df
		$\tau/\sigma = 1/4$	
Sign	1.49	1.57	1.88
Wilcoxon	1.76	1.83	2.21
Brown	1.86	1.93	2.34
Noether	2.12	2.14	2.48
Adaptive	2.12	2.14	2.48
		$\tau/\sigma = 1/2$	
Sign	2.24	2.48	3.44
Wilcoxon	3.17	3.40	4.74
Brown	3.60	3.83	5.39
Noether	4.97	4.72	5.77
Adaptive	4.97	4.72	5.77
		$\tau/\sigma = 3/4$	
Sign	3.41	3.90	6.02
Wilcoxon	5.92	6.42	9.70
Brown	7.55	7.91	11.69
Noether	13.48	10.86	12.08
Adaptive	13.48	10.86	12.08

or a central t-distribution with 3 degrees of freedom. For example, if one takes $\sigma = 1$, then for the Normal $\tau/\sigma = 1/2$ if $\tau = 1/2$, for the logistic $\tau/\sigma = 1/2$ if $\tau = (1/2)(\pi/\sqrt{3}) \doteq 0.907$, and for the t-distribution with 3 degrees of freedom, $\tau/\sigma = 1/2$ if $\tau = (1/2)\sqrt{3} \doteq 0.866$. Although $\widetilde{\Gamma}_{\text{no}} > \widetilde{\Gamma}_{\text{bmh}}$ throughout Table 2, there are many situations with $\widetilde{\Gamma}_{\text{no}} < \widetilde{\Gamma}_{\text{bmh}}$; for instance, with $\lambda_1 = 1/3$ this can occur in a t-distribution with 2 or 1 degrees of freedom, where the t with 1 degree of freedom is the Cauchy distribution, and it occurs in the t-distribution with 3 degrees of freedom in Table 6 of Section 6.1 with $\lambda_1 < 1/3$.

As an illustration of the properties of design sensitivity, consider the case of $\tau/\sigma = 1/2$ in Table 2 for the t-distribution with 3 degrees of freedom. The design sensitivity for Wilcoxon's statistic in this case is $\widetilde{\Gamma} = 4.74$, whereas for Noether's statistic it is $\widetilde{\Gamma}_{\text{no}} = 5.77$. In sufficiently large samples from this distribution, Wilcoxon's statistic should be sensitive to a bias of magnitude $\Gamma = 5$ but Noether's statistic should not. Drawing a single sample of $I = 10{,}000$ pairs from this distribution and performing a sensitivity analysis with $\Gamma = 5$ yields an upper bound on the P-value for Wilcoxon's statistic of 0.9985 and for Noether's statistic of 0.0071, so a deviation from random assignment of magnitude $\Gamma = 5$ could readily explain the observed value of Wilcoxon's statistic, but not the observed value of Noether's statistic. At the $\alpha = 0.011$ level with $\Gamma = 5$, the adaptive test rejects H_0 because Noether's statistic has passed its critical point in (4) although Brown's statistic has not. Because I was very large in this illustration, test performance was predicted by the design sensitivity, but in smaller sample sizes, both design sensitivity and efficiency affect test performance.

4.4. *Simulation*: *Power of a sensitivity analysis in the favorable situation*. The power of a sensitivity analysis is examined for finite I by simulation in Tables 3 and 4. The tables describe the favorable situation: there is a treatment effect and no bias from unobserved covariates, but of course the investigator does not know this in an observational study, and so performs a sensitivity analysis. The power of a 0.05-level sensitivity analysis is the probability that the upper bound on the one-sided P-value is less than 0.05. The power is determined for Wilcoxon's signed rank test, Brown's test, Noether's test and the adaptive test that uses both Brown's and Noether's tests.

In Tables 3 and 4, there is an additive effect and no bias from unobserved covariates, that is, $Y_i = \tau + \varepsilon_i$ and the ε_i are independent and identically distributed with a Normal, a logistic or a central t-distribution with 3 degrees of freedom. In Table 3, the effect is half the standard deviation σ of the ε_i's, $\tau/\sigma = 1/2$, whereas in Table 4, the effect is either $\tau/\sigma = 1/4$ or $\tau/\sigma = 3/4$.

Each sampling situation is replicated 10,000 times. Therefore, the standard error of the simulated power is at most $\sqrt{1/(4 \times 10{,}000)} = 0.005$. In each sampling situation for each Γ, the two highest powers are in *bold*.

Based on Table 1, we expect Wilcoxon's statistic to have the highest power for $\Gamma = 1$. Based on Table 2, we expect that for sufficiently large Γ and I, Noether's

TABLE 3

Simulated power with I pairs of a 0.05 *level sensitivity analysis performed with sensitivity parameter* Γ. *In each situation, there is no bias and there is an additive constant treatment effect* τ *whose magnitude is half the standard deviation of the pair differences* Y_i, *so* $\tau/\sigma = 1/2$. *Each situation is replicated* 10,000 *times. In each comparison, the two highest powers are in* bold

Pairs:	$I = 100$			$I = 250$			$I = 500$		
Γ:	**1**	**2**	**3**	**2**	**3**	**4**	**3**	**4**	**5**
				Normal errors, $\tau/\sigma = 1/2$					
Wilcoxon	**1.00**	0.53	0.05	0.89	0.07	0.00	0.10	0.00	0.00
Brown	**1.00**	0.61	0.10	0.93	0.19	0.01	0.35	0.01	0.00
Noether	0.99	**0.64**	**0.29**	**0.96**	**0.55**	**0.15**	**0.81**	**0.24**	**0.03**
Adaptive	1.00	**0.68**	**0.17**	**0.97**	**0.45**	**0.15**	**0.76**	**0.24**	**0.03**
				Logistic errors, $\tau/\sigma = 1/2$					
Wilcoxon	**1.00**	0.65	0.10	0.95	0.16	0.00	0.25	0.00	0.00
Brown	**1.00**	**0.70**	0.15	**0.97**	0.28	0.02	0.53	0.03	0.00
Noether	0.99	0.61	**0.26**	0.95	**0.48**	**0.11**	**0.75**	**0.17**	**0.02**
Adaptive	1.00	**0.70**	**0.18**	**0.97**	**0.41**	**0.11**	**0.70**	**0.17**	**0.02**
				t errors with 3 d.f., $\tau/\sigma = 1/2$					
Wilcoxon	**1.00**	**0.94**	0.45	**1.00**	0.81	0.21	0.98	0.33	0.02
Brown	**1.00**	**0.94**	**0.48**	**1.00**	**0.86**	**0.37**	**0.99**	**0.62**	0.10
Noether	1.00	0.77	0.42	0.99	0.75	0.29	0.95	0.50	**0.13**
Adaptive	1.00	0.92	**0.49**	1.00	**0.87**	**0.37**	**0.99**	**0.58**	**0.14**

statistic will have the highest power. Combining Tables 1 and 2, we see that, for the t-distribution with 3 degrees of freedom, Brown's statistic is much more efficient than Noether's statistic but has only slightly inferior design sensitivity, so Brown's statistic could have higher power for quite large I. Proposition 1 suggests that the adaptive procedure has fulfilled its potential if it has power close to the maximum of the powers of Brown's and Noether's statistics. With a few exceptions, these expectations are confirmed in Tables 3 and 4. Notably, in Tables 3 and 4, the adaptive procedure is never very bad, whereas other statistics perform poorly in some cases; for instance, in Table 4 the power loss is 90% for Wilcoxon's statistic compared to Noether's statistic for $I = 500$ pairs, Normal errors, $\tau/\sigma = 3/4$.

It is useful to contrast Tables 2, 3 and 4. For instance, the design sensitivity (as $I \to \infty$) of Noether's statistic is $\widetilde{\Gamma} = 4.97$ for matched pair differences Y_i that are $Y_i \sim_{\text{i.i.d.}} N(\frac{1}{2}, 1)$ in Table 2, but the power is only 15% in this case at $\Gamma = 4 < 4.97 = \widetilde{\Gamma}$ for $I = 250$ pairs in Table 3. That is, if $Y_i \sim_{\text{i.i.d.}} N(\frac{1}{2}, 1)$ with $I = 250$ pairs, there is an 89% chance the results will be sensitive at $\Gamma = 4$, even though as $I \to \infty$ the same distribution would eventually be seen to be insensitive at $\Gamma = 4$. The design sensitivity $\widetilde{\Gamma}$ in Table 2 refers to the limit as $I \to \infty$, so results will typically become sensitive at a smaller Γ, $\Gamma < \widetilde{\Gamma}$, in a finite sample,

TABLE 4

Simulated power with I pairs of a 0.05 *level sensitivity analysis performed with sensitivity parameter* Γ. *In each situation, there is no bias and there is an additive constant treatment effect* τ *whose magnitude is either* 1/4 *or* 3/4 *of the standard deviation* σ *of the pair differences* Y_i, *so* $\tau/\sigma = 1/4$ *or* $\tau/\sigma = 3/4$. *Each situation is replicated* 10,000 *times. In each comparison, the two highest powers are in* bold

	$\tau/\sigma = 1/4$				$\tau/\sigma = 3/4$			
Pairs:	$I = 100$		$I = 500$		$I = 100$		$I = 500$	
Γ:	**1**	**1.5**	**1.5**	**1.75**	**2.5**	**3.5**	**5**	**6**
				Normal errors				
Wilcoxon	**0.78**	0.16	0.44	0.05	0.92	0.49	0.28	0.02
Brown	**0.75**	0.18	0.53	0.11	0.94	0.67	0.78	0.33
Noether	0.60	**0.20**	**0.65**	**0.33**	**0.97**	**0.78**	**0.99**	**0.92**
Adaptive	0.72	**0.22**	**0.67**	**0.28**	**0.96**	**0.80**	**0.99**	**0.87**
				Logistic errors				
Wilcoxon	**0.83**	0.20	0.59	0.11	0.95	0.60	0.51	0.08
Brown	**0.79**	**0.21**	**0.66**	0.18	**0.95**	**0.71**	0.85	0.42
Noether	0.60	0.19	0.65	**0.33**	0.93	0.67	**0.93**	**0.75**
Adaptive	0.76	**0.23**	**0.70**	**0.29**	**0.96**	**0.73**	**0.94**	**0.68**
				t errors, 3 d.f.				
Wilcoxon	**0.96**	**0.47**	**0.97**	0.67	**1.00**	**0.92**	1.00	0.91
Brown	**0.94**	**0.47**	**0.98**	**0.74**	1.00	**0.93**	**1.00**	**0.97**
Noether	0.75	0.33	0.90	0.67	0.95	0.74	0.97	0.87
Adaptive	0.92	0.44	0.97	**0.74**	1.00	0.90	**1.00**	**0.97**

$I < \infty$. Although Noether's statistic is always best in Table 2, it is not always best in Tables 3 and 4; however, the adaptive test is never far behind the best test in Tables 3 and 4.

4.5. *Ties.* Ties are addressed in a straightforward manner when testing H_τ. Providing fewer than $(1 - \lambda_2)I$ of the $Y_i - \tau$ are equal to zero, no adjustment for zero differences is needed in the discussion in Section 4.1; that is, Brown's statistic and the adaptive procedure require no adjustment unless more than 1/3 of the sample is tied at zero. For ties among the $|Y_i - \tau|$, use average ranks in computing q_i; then I_1 and I_2 are random variables that depend upon the pattern of ties, but the procedure in Section 4.1 yields a test that is conditionally distribution-free given the realized values of I_1 and I_2.

5. Use of the adaptive procedure in the study of treatments for ovarian cancer. The matched pair difference in weeks with toxicity in the first year after

diagnosis is highly significant in randomization tests; for instance, the randomization based P-value from Wilcoxon's signed rank test is less than 10^{-6}. For $\Gamma = 1.3$ and $\Gamma = 1.6$, the upper bounds on the one-sided P-value from Wilcoxon's test are, respectively, 0.0032 and 0.128, so a bias of $\Gamma = 1.3$ could not easily produce the observed value of Wilcoxon's statistic, but a bias of $\Gamma = 1.6$ could do so. In contrast, the upper bound on the one-sided P-value from the adaptive procedure is 0.004 for $\Gamma = 1.6$, so the magnitude of bias that would explain the behavior of Wilcoxon's statistic does not begin to explain the behavior of the adaptive test. [The P-value at $\Gamma = 1.6$ for the adaptive test is the smallest α in (4)–(6) that leads to rejection.] The upper bound on the P-value from the adaptive test crosses 0.05 between $\Gamma = 1.96$ to $\Gamma = 1.97$. At $\Gamma = 1.96$, the adaptive procedure rejects based on Noether's statistic, which if used on its own would have an upper-bound on its one-sided P-value of 0.038. To put these quantities in context using the approach in Rosenbaum and Silber (2009), $\Gamma = 2$ corresponds with an unobserved covariate u_{ij} that produces a three-fold increase in the odds of greater toxicity and a five-fold increase in the odds of treatment by a medical oncologist, so the adaptive test reports considerably less sensitivity to bias from an unmeasured covariate than does Wilcoxon's test.

Similar results are found over the first five years. The upper bound on the P-value from Wilcoxon's test is 0.080 for $\Gamma = 1.7$, whereas for the adaptive test, the upper bound on the P-value is 0.047 for $\Gamma = 2.2$. As before, it is Noether's test, not Brown's test, that leads the adaptive test to reject.

To illustrate the calculations for toxicity in the first year, allowing for ties, Noether's statistic looks at the largest $I_1 = 100$ of the $|Y_i|$ finding $B_1 = 82$ of these have $Y_i > 0$, whereas Brown's statistic looks at the largest $I_1 + I_2 = 110 + 106 = 226$ of the $|Y_i|$ and the statistic has value $T = 2B_1 + B_2 = 222$. Using the binomial distribution as discussed in Section 4.1 with $\Gamma = 1.96$, one finds $\Pr(\overline{\overline{B}}_1 \geq 82) = 0.0381$, $\Pr(\overline{\overline{T}}_1 \geq 237) = 0.0293$, $\Pr(\overline{\overline{B}}_1 \geq 82 \text{ or } \overline{\overline{T}}_1 \geq 237) = 0.0475$, so the adaptive procedure rejects at the 0.05 level for every bias less than $\Gamma = 1.96$, but only Noether's test, not Brown's test, would have led to rejection used on its own.

In Section 6.1 the choice of (λ_1, λ_2) is discussed. As a prelude to that discussion, consider the results of the sensitivity analysis for toxicity in the first year for two choices of (λ_1, λ_2) besides $(1/3, 2/3)$. If $\lambda_1 = 1/6$ and $\lambda_2 = 2/6$ are used in place of $\lambda_1 = 1/3$ and $\lambda_2 = 2/3$, the adaptive procedure has an upper bound on the one-sided P-value of 0.046 for $\Gamma = 3.3$. If $\lambda_1 = 1/8$ and $\lambda_2 = 2/8$ are used, the adaptive procedure has an upper bound on the one-sided P-value of 0.046 for $\Gamma = 3.7$. Using $\lambda_1 = 1/8$ and allowing for ties in the $I = 344$ pairs, Noether's statistic focuses on the 45 of 344 pairs with the largest $|Y_i|$ and finds that 41 of these 45 pairs have $Y_i > 0$. In words, when there was a large difference in weeks with toxicity, it was usually the result of greater toxicity in a patient treated by a medical oncologist, and this seems unlikely to have occurred by chance if the magnitude of bias from nonrandom assignment is $\Gamma \leq 3.7$.

6. Discussion.

6.1. *Variations on a theme*: *Other λ's*; *other statistics*. The adaptive procedure in Section 4 uses two compatible tests statistics from the statistical literature. Brown's (1981) statistic was designed to be a serious competitor of Wilcoxon's statistic in a randomized experiment without bias, yet Brown's statistic has higher design sensitivity when errors are Normal or logistic or t-distributed with 3 degrees of freedom. The version of Noether's (1973) test used here has poor Pitman efficiency in these cases but much better design sensitivity. So the adaptive procedure adapts between a procedure with good Pitman efficiency with good design sensitivity and a procedure with poor Pitman efficiency and excellent design sensitivity. There are, of course, many possible variations on this theme, some more promising than others.

The statistics of Brown and Noether take one or two large steps, but otherwise are constant as functions of the ranks q_i of the $|Y_i|$. Are large flat steps useful? Both statistics decrease the weight attached to small $|Y_i|$ and increase the weight attached to large $|Y_i|$ without emphasizing the extremely large $|Y_i|$. Would a gradual increase be better than a step? Consider ranks that equal q_i/I if $q_i/I \geq 1-\lambda$ and equal 0 if $q_i/I < 1-\lambda$; call this the "$(1-\lambda)$-step Wilcoxon statistic" because it uses Wilcoxon's ranks above $1-\lambda$. Wilcoxon's statistic is the 0-step Wilcoxon statistic. The $2/3$-step Wilcoxon statistic takes a step where Noether's statistic takes a step, but it increases gradually thereafter, and the $1/3$-step Wilcoxon statistic takes a step where Brown's statistic takes its first step, but it increases gradually thereafter. Table 5 contrasts the design sensitivities of Brown's statistic, Noether's statistic and comparable step-Wilcoxon statistics in the case of an additive treatment effect whose magnitude is half the standard deviation of the errors. While the difference between Brown's statistic and Noether's statistic is large, the difference between either of these and its comparable step-Wilcoxon statistic is not large.

Brown's statistic focuses on the largest $2/3$ of the $|Y_i|$, while Noether's statistic focuses on the largest $1/3$ of $|Y_i|$. In the example in Section 1.2, further tinkering

TABLE 5

Design sensitivities for Brown's statistic, Noether's statistic and for two comparable step-Wilcoxon statistics. The table refers to an additive treatment effect that is half the standard deviation of the errors, for errors with a Normal distribution, a logistic distribution or a t-distribution with 3 *degrees of freedom*

	Normal	**Logistic**	***t* 3 df**
Brown	3.60	3.83	5.39
1/3-step Wilcoxon	3.60	3.83	5.35
Noether	4.97	4.72	5.77
2/3-step Wilcoxon	5.20	4.80	5.64

TABLE 6

Design sensitivities for the Brown–Markowski–Hettmansperger statistic, Noether's statistic and the adaptive statistic for various values of λ_1 *with* $\lambda_2 = 2\lambda_1$. *The table refers to an additive treatment effect that is half the standard deviation of the errors, for errors with a Normal distribution, a logistic distribution or a t-distribution with* 3 *degrees of freedom. The largest design sensitivity in a sampling situation (or in a column) is in* bold

	λ_1	Normal	Logistic	t 3 df
Brown–Markowski–Hettmansperger	1/3	3.60	3.83	5.39
Noether	1/3	4.97	4.72	**5.77**
Adaptive	1/3	4.97	4.72	**5.77**
Brown–Markowski–Hettmansperger	1/4	4.36	4.37	5.67
Noether	1/4	5.87	5.06	5.53
Adaptive	1/4	5.87	5.06	5.67
Brown–Markowski–Hettmansperger	1/6	5.58	4.93	5.51
Noether	1/6	7.28	5.41	5.03
Adaptive	1/6	7.28	5.41	5.51
Brown–Markowski–Hettmansperger	1/8	6.55	5.23	5.20
Noether	1/8	**8.40**	**5.59**	4.64
Adaptive	1/8	**8.40**	**5.59**	5.20

with λ_1 and λ_2 led to greater insensitivity to unmeasured bias. Markowski and Hettmansperger (1982) discuss the choice of λ_1 and λ_2 from the perspective of Pitman efficiency. Table 6 compares the design sensitivities for several values of λ_1 with $\lambda_2 = 2\lambda_1$.

In thinking about Table 6, several cautions are needed. First, the design sensitivity refers to a limit as the number I of pairs increases $I \to \infty$, so Table 6 is unlikely to offer useful guidance unless $I\lambda_1$ is a reasonably large number. With $I = 100$ pairs and $\lambda_1 = 1/8$, there are only 13 pairs counted in Noether's statistic, so asymptotic theory is not likely to provide useful guidance. Second, the Pitman efficiencies for Noether's statistic with $\lambda_1 < 1/3$ are substantially worse than the already disappointing values shown in Table 1, so Table 6 is only relevant when the sample size I is so large that the design sensitivity has come to dominate the Pitman efficiency, as it will do in the limit as $I \to \infty$ because the power function tends to a step function dropping from power 1 to power 0 at $\widetilde{\Gamma}$. There are, however, many large observational studies, for example, Volpp et al. (2007) conducted an observational study of 8.5 million hospital admissions. Third, the columns of Table 6 refer to distributions that differ greatly in their tails, so an answer that depends strongly upon which column is considered is an answer that depends strongly on the behavior of the most extreme observations.

With these cautions firmly in mind, consider Table 6. In Table 6, $\lambda_1 = 1/8$ is best for the Normal and logistic distributions and $\lambda_1 = 1/3$ is best for the t-distribution with 3 degrees of freedom; see Rosenbaum [(2010a), Figure 2] for a heuristic ex-

planation of the relationship between tail behavior, weights and sensitivity to bias. With smaller λ_1's, the adaptive procedure looks attractive: for $\lambda_1 = 1/8$ it uses Noether's test to advantage for Normal errors and it uses the Brown–Markowski–Hettmansperger test for t-errors. Notably, in Table 6, the adaptive procedure exhibits relatively stable performance as λ_1 decreases for the t-distribution, but it captures large gains for the Normal and logistic distributions.

6.2. *Are large observational studies less susceptible to unmeasured biases*? Section 1 began with the question: are large observational studies less susceptible to unmeasured biases? The success of the adaptive procedure suggests that this question is incorrectly posed. An observational study is sensitive to biases of a certain magnitude, and the sample size is not the key element in determining this. However, a poor choice of test statistic—perhaps the Wilcoxon statistic—may lead to a sensitivity analysis that exaggerates the degree of sensitivity to unmeasured biases. A good choice of test statistic may depend upon features of the observable distributions that are unknown to the investigator prior to the investigation. To the extent that a large sample size permits us to see clearly these features of observable distributions, it may let us adapt the statistical analysis so that a poor choice of test statistic does not exaggerate the degree of sensitivity to unmeasured biases.

REFERENCES

BROWN, B. M. (1981). Symmetric quantile averages and related estimators. *Biometrika* **68** 235–242. MR0614960

COCHRAN, W. G. (1965). The planning of observational studies of human populations (with discussion). *J. Roy. Statist. Soc. Ser. A* **128** 234–266.

COPAS, J. and EGUCHI, S. (2001). Local sensitivity approximations for selectivity bias. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63** 871–895. MR1872072

CORNFIELD, J., HAENSZEL, W., HAMMOND, E. C., LILIENFELD, A. M., SHIMKIN, M. B. and WYNDER, E. L. (1959). Smoking and lung cancer. *J. Natl. Cancer Inst.* **22** 173–203.

COX, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444. MR0378189

DIPRETE, T. A. and GANGL, M. (2004). Assessing bias in the estimation of causal effects. *Sociol. Method.* **34** 271–310.

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. MR2751473

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

FRANGAKIS, C. E. and RUBIN, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika* **86** 365–379. MR1705410

GADBURY, G. L. (2001). Randomization inference and bias of standard errors. *Amer. Statist.* **55** 310–313. MR1939365

GASTWIRTH, J. L. (1966). On robust procedures. *J. Amer. Statist. Assoc.* **61** 929–948. MR0205397

GASTWIRTH, J. L. (1992). Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics* **33** 19–34.

GILBERT, P. B., BOSCH, R. J. and HUDGENS, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics* **59** 531–541. MR2004258

GROENEVELD, R. A. (1972). Asymptotically optimal group rank tests for location. *J. Amer. Statist. Assoc.* **67** 847–849.

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. (1963). Estimates of location based on rank tests. *Ann. Math. Statist.* **34** 598–611. MR0152070

HOGG, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory (with discussion). *J. Amer. Statist. Assoc.* **69** 909–923.

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

IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93** 126–132.

JONES, D. H. (1979). An efficient adaptive distribution-free test for location. *J. Amer. Statist. Assoc.* **74** 822–828. MR0556475

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

LIN, D. Y., PSATY, B. M. and KRONMAL, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54** 948–963.

MARCUS, S. M. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *J. Educ. Behav. Statist.* **22** 193–201.

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

MARKOWSKI, E. P. and HETTMANSPERGER, T. P. (1982). Inference based on simple rank step score statistics for the location model. *J. Amer. Statist. Assoc.* **77** 901–907. MR0686416

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

NEYMAN, J. (1923). On the application of probability theory to agricultural experiments. *Statist. Sci.* **5** 463–480.

NOETHER, G. (1973). Some distribution-free confidence intervals for the center of a symmetric distribution. *J. Amer. Statist. Assoc.* **68** 716–719.

POLICELLO, G. E. and HETTMANSPERGER, T. P. (1976). Adaptive robust procedures for the one-sample location problem. *J. Amer. Statist. Assoc.* **71** 624–633.

REITER, J. (2000). Using statistics to determine causal relationships. *Amer. Math. Monthly* **107** 24–32. MR1543589

ROSENBAUM, P. R. (1993). Hodges–Lehmann point estimates of treatment effect in observational studies. *J. Amer. Statist. Assoc.* **88** 1250–1253. MR1245357

ROSENBAUM, P. R. (2002a). *Observational Studies*, 2nd ed. Springer, New York.

ROSENBAUM, P. R. (2002b). Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.* **17** 286–327. MR1962487

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

ROSENBAUM, P. R. (2005). Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies. *Amer. Statist.* **59** 147–152. MR2133562

ROSENBAUM, P. R. (2010a). Design sensitivity and efficiency in observational studies. *J. Amer. Statist. Assoc.* **105** 692–702. MR2724853

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

ROSENBAUM, P. R. (2011). A new u-statistic with superior design sensitivity in observational studies. *Biometrics* **67** 1017–1027.

ROSENBAUM, P. R., ROSS, R. N. and SILBER, J. H. (2007). Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Amer. Statist. Assoc.* **102** 75–83. MR2345534

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

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

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

RUBIN, D. B. (1979). Using multivariate matched sampling and regression adjustment to control bias in observational studies. *J. Amer. Statist. Assoc.* **74** 318–328.

SILBER, J. H., ROSENBAUM, P. R., POLSKY, D., ROSS, R. N., EVEN-SHOSHAN, O., SCHWARTZ, S., ARMSTRONG, K. A. and RANDALL, T. C. (2007). Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? *J. Clin. Oncol.* **25** 1169–1175. Related editorial: **25** 1157–1158. Related letters and rejoinders: **25** 3551–3558.

SMALL, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J. Amer. Statist. Assoc.* **102** 1049–1058. MR2411664

VOLPP, K. G., ROSEN, A. K., ROSENBAUM, P. R., ROMANO, P. S., EVEN-SHOSHAN, O., WANG, Y., BELLINI, L., BEHRINGER, T. and SILBER, J. H. (2007). Mortality among hospitalized Medicare beneficiaries in the first 2 years following ACGME resident duty hour reform. *J. Am. Med. Assoc.* **298** 975–983.

WANG, L. and KRIEGER, A. M. (2006). Causal conclusions are most sensitive to unobserved binary covariates. *Stat. Med.* **25** 2257–2271. MR2240099

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

WILK, M. B. and GNANADESIKAN, R. (1968). Probability plotting methods for the analysis of data. *Biometrika* **55** 1–17.

WOLFE, D. A. (1974). A characterization of population weighted-symmetry and related results. *J. Amer. Statist. Assoc.* **69** 819–822. MR0426239

YANAGAWA, T. (1984). Case-control studies: Assessing the effect of a confounding factor. *Biometrika* **71** 191–194. MR0738341

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

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
473 JON M. HUNTSMAN HALL
3730 WALNUT STREET
PHILADELPHIA, PENNSYLVANIA 19104-6340
USA
E-MAIL: rosenbaum@wharton.upenn.edu