Published in final edited form as:
J Am Stat Assoc. 2009 December 1; 104(488): 1398–1405. doi:10.1198/jasa.2009.tm08470.

Amplification of Sensitivity Analysis in Matched Observational Studies

Paul R. Rosenbaum and **Jeffrey H. Silber**[1]
University of Pennsylvania, Philadelphia

Abstract

A sensitivity analysis displays the increase in uncertainty that attends an inference when a key assumption is relaxed. In matched observational studies of treatment effects, a key assumption in some analyses is that subjects matched for observed covariates are comparable, and this assumption is relaxed by positing a relevant covariate that was not observed and not controlled by matching. What properties would such an unobserved covariate need to have to materially alter the inference about treatment effects? For ease of calculation and reporting, it is convenient that the sensitivity analysis be of low dimension, perhaps indexed by a scalar sensitivity parameter, but for interpretation in specific contexts, a higher dimensional analysis may be of greater relevance. An amplification of a sensitivity analysis is defined as a map from each point in a low dimensional sensitivity analysis to a set of points, perhaps a 'curve,' in a higher dimensional sensitivity analysis such that the possible inferences are the same for all points in the set. Possessing an amplification, an investigator may calculate and report the low dimensional analysis, yet have available the interpretations of the higher dimensional analysis.

Keywords

Amplification; causal effects; observational study; sensitivity analysis

1 Example: Is More Chemotherapy More Effective, or Just More Toxic?

In discussing what credence to give to an observational or nonrandomized study of treatment effects, a central concern is that adjustments for measured covariates may fail to render the treated and control groups comparable prior to treatment, so that differing outcomes in treated and control groups may not be effects caused by the treatment. In this context, a sensitivity analysis asks what attributes an unobserved covariate would have to have to materially alter the conclusions.

Our purpose in the current paper is to discuss a new device we call 'amplification' which can aid in interpreting the results of a sensitivity analysis without increasing the complexity of the sensitivity analysis or the space in an empirical article needed to display it. By mapping a one-dimensional sensitivity analysis into a higher dimensional sensitivity analysis, the interpretations of the higher dimensional analysis are available when only the one dimensional analysis is performed and reported. To illustrate these ideas, we use a study by Silber, et al. (2007), and a brief summary of that study follows. The study asked whether greater intensity of chemotherapy for ovarian cancer was beneficial to patients in terms of survival, or whether it simply increased toxicity.

[1]*Address for correspondence*: Department of Statistics, The Wharton School, University of Pennsylvania, 473 Jon M. Huntsman Hall, 3730 Walnut Street, Philadelphia, PA 19104-6340 USA. rosenbaum@stat.wharton.upenn.edu.

Outside of clinical trials, where treatments are assigned at random, much of the variation in the treatments that patients receive is a direct response to variations in the health of those patients. The intensity of the chemotherapy may vary from patient to patient in response to variations in the health of patients prior to treatment. Random treatment assignment is one way, the best way, to create variation in treatment that is not a response to variation in the health of patients, but this best source of unconfounded variation in treatment is not available outside of randomized clinical trials. Is there a source of variation in the intensity of chemotherapy that is not a reaction to variation in the health of patients?

Ovarian cancer is unusual in that two medical specialties provide chemotherapy, medical oncologists (MOs) and gynecologic oncologists (GOs). To the extent that MOs and GOs differ in the way they use chemotherapy, they may provide some variation in medical practice that is not primarily a response to the health of their patients. Medical oncologists have specialized training in the provision of chemotherapy for cancers of all kinds. Gynecologic oncologists are trained first in gynecology and then receive additional training in oncology, and they typically treat cancers of the ovary, uterus and cervix. As gynecologists, GOs are surgeons, and they often perform the surgery that precedes chemotherapy, whereas MOs are almost invariably not surgeons, so they provide chemotherapy after someone else has performed the surgery. Silber, et al. anticipated correctly that MOs would often use chemotherapy more intensively than GOs, and they sought to use this variation in treatment style to ask whether greater intensity was beneficial. Moreover, after surgery, the principal treatment for ovarian cancer is chemotherapy, so if MO's and GO's produce differing clinical outcomes, it is very likely to be the result of differences in their use of chemotherapy. The data merged U.S. Medicare claims and follow-up with clinical information from the Surveillance, Epidemiology and End Results program of the U.S. National Cancer Institute. Patients were classified as MO or GO patients based on the dominant provider of chemotherapy during the first three months after diagnosis. Using a matching procedure described in Rosenbaum, Ross and Silber (2007), they matched 344 patients of GOs to 344 similar patients of MOs, matching for 36 covariates, including clinical stage, tumor grade, surgeon type, a variety of comorbid conditions such as congestive heart failure and diabetes, demographic variables such age and race, SEER site, and year of diagnosis. As seen in Tables 2 and 3 of Silber, et al. (2007), after matching, the MO and GO patient groups were highly comparable on these 36 measured covariates, the balance being somewhat better than obtained by complete randomization, but of course in an observational study there may be systematic differences that were not measured. As seen in Figure 1 and Table 4 of Silber, et al. (2007), survival was virtually identical in the two groups, but MO's provided more weeks of chemotherapy and produced more weeks of reported chemotherapy related toxicity, such as anemia, neutropenia, thrombocytopenia, and drug induced neuropathy, both during initial treatment and in later years following cancer recurrence.

Here, we focus on intensity of initial treatment and toxicity during the first year after diagnosis. In particular, we focus on matched pair differences, MO-minus-GO, because the pairs are closely matched for important clinical variables, such as stage and grade and surgeon type. Figure 1 is a plot of 344 matched pair differences in toxicity weeks against 344 matched pair differences in chemotherapy weeks, so both differences can take integer values between −52 and 52. Figure 1 also displays the lowess smooth (using the defaults in R) and the marginal boxplot of the 344 matched pair differences in toxicity weeks. Incidentally, tinkering with the settings of the lowess smooth does not alter its qualitative appearance, except for introducing small wiggles. There are several notable patterns in Figure 1. First, there is a dense cloud of points near the origin, (0, 0), suggesting that in many matched pairs of two similar patients, the MOs and GOs produced similar toxicity with treatments of similar intensity, perhaps following regimes evaluated in the clinical trials

that usually guide initial treatment for cancer. In the upper right quadrant, however, a subset of patients treated by MOs received more weeks of chemotherapy than their matched GO patients and experienced substantially more toxicity. This might suggest the model of Conover and Salsburg (1988) in which a treatment (here, the MO vs GO distinction) affects only a subset of patients. What attributes would an unmeasured covariate have to possess to produce the difference in toxicity seen in Figure 1?

2 Notation and Review

2.1 Treatment Effects and Treatment Assignments

There are I matched pairs, $i = 1,\ldots, I$, of two subjects, $j = 1, 2$, one treated, denoted $Z_{ij} = 1$, the other control, denoted $Z_{ij} = 0$, matched for observed covariates, $\mathbf{x}_{ij}$, so that $1 = Z_{i1} + Z_{i2}$ and $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each i. There is concern about the possible impact of failing to match for a covariate u_{ij} that was not observed, where typically $u_{i1} \neq u_{i2}$. Following Neyman (1923) and Rubin (1974)'s notation for treatment effects, each subject ij has two potential responses, (r_{Tij}, r_{Cij}), and this subject would exhibit response r_{Cij} if assigned to control, $Z_{ij} = 0$, or response r_{Tij} if assigned to treatment, $Z_{ij} = 1$, so the response actually observed from this subject is $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})\, r_{Cij}$, and the effect of the treatment, namely $r_{Tij} - r_{Cij}$, is not observed for any subject ij. Write $\mathbf{Z} = (Z_{11},\ldots, Z_{I2})^T$, $\mathbf{R} = (R_{11},\ldots, R_{I2})^T$, $\mathbf{r}_C = (r_{C11},\ldots, r_{CI2})^T$, $\mathbf{r}_T = (r_{T11},\ldots, r_{TI2})^T$ and $\mathbf{u} = (u_{11},\ldots, u_{I2})^T$ for the $2I$-dimensional vectors. Fisher's sharp null hypothesis of no treatment effect asserts that the response of each subject ij is unchanged by receiving the treatment, $H_0 : r_{Tij} = r_{Cij}, \forall_{ij}$ or $H_0 : \mathbf{r}_C = \mathbf{r}_T$, and if the hypothesis is true then $\mathbf{R} = \mathbf{r}_C$.

Write $\mathcal{F} = \{(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}), i = 1,\ldots, I, j = 1, 2\}$, and write $\mathcal{Z}$ for the set containing the 2^I possible treatment assignments $\mathbf{Z}$, so that $\mathbf{z} \in \mathcal{Z}$ if each z_{ij} is 0 or 1 and $z_{i1} + z_{i2} = 1$ for each i. The number of elements in a set S is denoted $|S|$, so $|\mathcal{Z}| = 2^I$.

2.2 Randomization Inference in Randomized Experiments

In a randomized, paired experiment, randomization ensures that $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}) = 2^{-I}$ for each $\mathbf{z} \in \mathcal{Z}$ In Fisher's (1935) phrase, randomization forms the "reasoned basis for inference" in such an experiment, in the specific sense that the distribution of any test statistic, $t(\mathbf{Z}, \mathbf{R})$, under the null hypothesis of no effect, H_0, is its permutation distribution

$$\Pr\{t(\mathbf{Z}, \mathbf{R}) \geq k \mid \mathcal{F}\} = \Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq k \mid \mathcal{F}\} = \frac{|\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{z}, \mathbf{R}) \geq k\}|}{2^I}, \quad (1)$$

because under H_0, $\mathbf{R} = \mathbf{r}_C$ is fixed by conditioning on $\mathcal{F}$ and $\mathbf{Z}$ is uniformly distributed on $\mathcal{Z}$ Write $Y_{Ci} = r_{Ci1} - r_{Ci2}$, so $Y_i = Y_{Ci}$ when the null hypothesis, H_0, of no treatment effect is true.

The results discussed here apply to most of the commonly used statistics for matched pairs, including Wilcoxon's signed rank statistic, McNemar's test for paired binary responses, the permutational t-statistic, permutation distributions for Huber's m-estimates, and others. Some specifics follow. Let $V_i = Z_{i1} - Z_{i2}$, $Y_i = R_{i1} - R_{i2}$, $\mathbf{A} = (|Y_1|,\ldots, |Y_I|)^T$, and let $q_i = q_i(\mathbf{A}) \geq 0$ be a function of $\mathbf{A}$ for each i such that $|Y_i| = A_i = 0$ implies $q_i = q_i(\mathbf{A}) = 0$. Test statistics widely use for permutation inference in matched pairs are of the form $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} q_i \operatorname{sign}(V_i Y_i)$ or are linear functions of a such a statistic, where $\operatorname{sign}(w) = 1, 0$ or -1 as $w > 0$, $w = 0$, or $w < 0$. For instance, if $q_i = |Y_i|$, then $t(\mathbf{Z}, \mathbf{R}) / I$ is the mean treated-minus-control difference in responses, and (1) is equivalent to the permutational t-test (Fisher 1935, Welch 1937). In the absence of ties, if q_i is the rank of $|Y_i|$, then $t(\mathbf{Z}, \mathbf{R})/2 + I(I + 1)/4$ is Wilcoxon's signed rank statistic, and similar considerations apply to

Stephenson's (1981) generalization of Wilcoxon's statistic which is particularly Effective when only some subjects respond to treatment (Conover and Salsburg 1988, Rosenbaum 2007a).

Many other statistics can be put in the same form, with their randomization distributions developed in a parallel way. If $\psi(\cdot)$ is an odd function, $\psi(-y) = -\psi(y)$, then $\sum_{i=1}^{I}\psi(V_i Y_i) = \sum_{i=1}^{I} q_i \text{ sign}(V_i Y_i)$ for $q_i = \psi(|Y_i|)$ yields the permutation distribution of a test based on Huber's m-estimate; see Maritz (1979) and Rosenbaum (2007b). The sign test, Noether's (1973) quantile average test, and Brown's (1981) combined quantile average test also have this form. In the special case in which R_{ij} is binary, Y_i is 1, 0, or −1, with $q_i = |Y_i| = 0$ for concordant pairs, $q_i = |Y_i| = 1$ for discordant pairs, yielding McNemar's test, in which (1) gives $\{t(\mathbf{Z}, \mathbf{R}) + n\}/2$ a binomial null distribution with sample size $n = \sum q_i$ and probability of success $\frac{1}{2}$.

For all of these statistics, under the null hypothesis, H_0, both $Y_i = Y_{Ci} = r_{Ci1} - r_{Ci2}$ and q_i are fixed in (1) by conditioning on $\mathcal{F}$, and in a randomized experiment V_i is 1 or −1 each with probability $\frac{1}{2}$, so the distribution of $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} q_i \text{ sign}(V_i Y_i)$ in (1) is the sum of I independent random variables each having support on one point (if $q_i = 0$) or two equally probable points $\pm q_i$. This yields simple approximations to (1) using the central limit theorem. Alternatively, exact calculations may use the method of Pagano and Tritchler (1983).

2.3 Observational Studies and Sensitivity Analysis

In the absence of randomization, there may be little basis for believing that $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}) = 2^{-I}$. A sensitivity analysis considers departures from random assignment of various magnitudes and their impact on inferences about treatment effects. Let $\mathcal{U} = [0, 1]^{2I}$ be the $2I$-dimensional unit cube. A convenient one parameter family of departures from random assignment is

$$\Pr(\mathbf{Z}=\mathbf{z} \mid \mathcal{F}) = \frac{\exp(\gamma \mathbf{z}^T \mathbf{u})}{\sum_{\mathbf{b}\in\mathcal{Z}} \exp(\gamma \mathbf{b}^T \mathbf{u})} = \prod_{i=1}^{I} \frac{\exp\{\gamma(z_{i1}u_{i1} + z_{i2}u_{i2})\}}{\exp(\gamma u_{i1}) + \exp(\gamma u_{i2})}, \mathbf{u} \in \mathcal{U} \tag{2}$$

for $\mathbf{z} \in \mathcal{Z}$ see Rosenbaum (1987) for the case of matched pairs and Rosenbaum (2002, §4) for extensions to other cases. It is easy to show (Rosenbaum 2002, §4.2) that model (2) is equivalent to assuming the following model: (i) in the population prior to matching, treatment assignments are independent, with $\boldsymbol{\pi}_{ij} = \Pr(Z_{ij} = 1 \mid \mathcal{F})$ as the probability of treatment; (ii) two subjects, ij and ij', with the same observed covariates, $\mathbf{x}_{ij} = \mathbf{x}_{ij'}$, may differ in their odds of receiving the treatment by at most a factor of $\Gamma = \exp(\gamma) \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1-\pi_{ij'})}{\pi_{ij'}(1-\pi_{ij})} \leq \Gamma; \tag{3}$$

then, the distribution of $\mathbf{Z}$ is restricted to $\mathcal{Z}$ by conditioning on $Z_{i1} + Z_{i2} = 1, \forall i$; finally, in (2), $u_{ij} = \{\log(\boldsymbol{\pi}_{ij}) - \min_k \log(\boldsymbol{\pi}_{ik})\} / \boldsymbol{\gamma}$. Expressed in the form (3), the sensitivity analysis is similar in spirit to the first sensitivity analysis proposed by Cornfield, et al. (1959); see also Gastwirth (1992) and Wang and Kreiger (2006). Under the null hypothesis H_0 and (2), the distribution of $t(\mathbf{Z}, \mathbf{R}) = t(\mathbf{Z}, \mathbf{r}_C)$ is given by f

$$\Pr\{t(\mathbf{Z},\mathbf{r}_C)\geq k\mid\mathscr{F}\}=\sum_{\mathbf{z}\in\mathscr{Z}}\chi\{t(\mathbf{z},\mathbf{r}_C)\geq k\}\prod_{i=1}^{I}\frac{\exp\{\gamma(z_{i1}u_{i1}+z_{i2}u_{i2})\}}{\exp(\gamma u_{i1})+\exp(\gamma u_{i2})} \tag{4}$$

where $\chi(E)=1$ if the event E occurs and $\chi(E)=0$ otherwise. For the statistics $t(\mathbf{Z},\mathbf{R})=\sum_{i=1}^{I}q_i\,\text{sign}(V_iY_i)$ defined in §2.2, define $\overline{\overline{T}}_\Gamma$ to be the sum of I independent random variables taking value q_i with probability $\Gamma/(1+\Gamma)$ and value $-q_i$ with probability $(1+\Gamma)^{-1}$, where $\Gamma=\exp(\gamma)$. In parallel, define $\overline{T}_\Gamma$ to be the sum of I independent random variables taking values q_i with probability $(1+\Gamma)^{-1}$ and value $-q_i$ with probability $\Gamma/(1+\Gamma)$. Notice that $\overline{T}_{1/\Gamma}$ and $\overline{\overline{T}}_\Gamma$ have the same distribution. It is straightforward to show (e.g., Rosenbaum 1987; 2002, §4.3) that under H_0 and (2) with $\Gamma=\exp(\gamma)\geq 1$,

$$\Pr(\overline{T}_\Gamma\geq k)\leq\Pr\{t(\mathbf{Z},\mathbf{r}_C)\geq k\mid\mathscr{F}\}\leq\Pr\left(\overline{\overline{T}}_\Gamma\geq k\right)\ \text{for all}\ \mathbf{u}\in\mathscr{U}, \tag{5}$$

where, again, $\mathbf{R}=\mathbf{r}_C$ when H_0 is true. Of course, the inequalities in (5) are reversed for $\Gamma=\exp(\gamma)\leq 1$. Under H_0, with $\Gamma=1$, the bounds in (5) are equal $\Pr(\overline{T}_1\geq k)=\Pr(\overline{\overline{T}}_1\geq k)$ and are equal to the randomization distribution (1), but as Γ increases, there is greater uncertainty about the correct null distribution for $t(\mathbf{Z},\mathbf{R})$. For each fixed $\Gamma\geq 1$, (5) yields an interval of possible significance levels, and by inversion, an interval of point estimates or an interval of endpoints for confidence intervals. The bounds in (5) are sharp in the sense that they are attained for particular covariates $\mathbf{u}\in\mathscr{U}$ so the bounds cannot be narrowed unless additional information is provided about $\mathbf{u}$. Specifically, the upper bound in (5) is attained for a $\mathbf{u}\in\mathscr{U}$ with $|u_{i1}-u_{i2}|=1$ and $\text{sign}(u_{i1}-u_{i2})=\text{sign}(r_{Ci1}-r_{Ci2})$ for all i, so u_{ij} is both as imbalanced as it can be and also as strongly associated with $r_{Ci1}-r_{Ci2}$ as it can be. The amplification in §3 will describe the same bounds (5) in terms of unobserved covariates $\mathbf{u}$ with very different properties.

Exact calculation of the bound (5) is feasible (see Rosenbaum 2003, Appendix, for software); however, for moderately large I, an approximation based on the central limit theorem is convenient and adequate. In particular, $\Pr(\overline{\overline{T}}_\Gamma\geq k)$ is approximated by

$$\Pr\left(\overline{\overline{T}}_\Gamma\geq k\right)\approx 1-\Phi\left\{\frac{(k-\frac{\Gamma-1}{\Gamma+1}\sum q_i)}{\sqrt{\frac{4\Gamma}{(1+\Gamma)^2}\sum q_i^2}}\right\},$$

where $\Phi(\cdot)$ is the standard Normal cumulative distribution.

Alternative methods of sensitivity analysis are discussed by Cornfield, et al. (1959), Rosenbaum and Rubin (1983), Yanagawa (1984), Gastwirth, et al. (1992, 1999), Robins, et al. (1999), Copas and Eguchi (2001), and Imbens (2003). For several applications, see Aakvik (2001), Diprete and Gangl (2004), Silber, et al. (2005), and Slade, et al. (2008).

3 Amplification of a one-dimensional sensitivity analysis

3.1 Unobserved covariate with a limited relationship with both treatment and response

The one parameter sensitivity model (2) limited the strength of the association between u_{ij} and Z_{ij} using the parameter $\Gamma=\exp(\gamma)$, and the bounds in (5) were attained for a u_{ij} with an extremely strong, near perfect, relationship with r_{Cij}. This bounding covariate may not be the one of greatest concern in a particular observational study, because a near perfect relationship between u_{ij} and r_{Cij} may be implausible. The alternative model considered in

this section has two parameters, $\Lambda = \exp(\lambda)$ controlling the strength of the relationship between u_{ij} and Z_{ij}, and $\Delta = \exp(\delta)$ controlling the relationship between u_{ij} and r_{Cij}. Because u_{ij} is related to both Z_{ij} and r_{Cij}, it can lead Z_{ij} and r_{Cij} to be associated in the absence of adjustment for u_{ij}. The amplification will map a value of Γ into a set or curve of values of (Λ, Δ) such that the one-dimensional analysis for Γ in §2.3 is correct for the entire curve of values of (Λ, Δ). Therefore, an investigator may perform and display the one-dimensional analysis in terms of Γ, yet interpret that analysis in terms of the two-dimensional (Λ, Δ) without further calculation and display. The parameter $\Delta = \exp(\delta)$ is the scalar parameter in a semiparametric family of distributions introduced by Wolfe (1974), which is briefly reviewed in §3.2.

Gastwirth, Krieger and Rosenbaum (GKR 1998) proposed a two-parameter sensitivity analysis involving natural parameter exponential family distributions, in which the one parameter analysis in §2.3 is the limiting case as the other parameter increases to infinity. However, except in the case of binary responses, the GKR analysis does not permit an amplification, in the sense that finite values of the two parameters do not correspond with values of Γ in §2.3, so one must report a two-dimensional array of statistical analysis, perhaps an array of upper bounds on significance levels or an array of intervals of point estimates. In contrast, the model in §3.3 permits a one-dimensional sensitivity analysis to be given a two-dimensional interpretation. The two approaches agree for binary responses using McNemar's test, but not in other cases.

3.2 Wolfe's semiparametric family of asymmetric distributions

Wolfe (1974) introduced a semiparametric family of asymmetric distributions in which a scalar parameter $\Delta > 0$ deforms an arbitrary symmetric distribution. Let W be a random variable symmetric about zero, so that $\Pr(W \leq -w) = \Pr(W \geq w)$ for each $w > 0$; here, W may be continuous or discrete or some combination of the two. Let W_Δ be a random variable such that $\Pr(W_\Delta = 0) = \Pr(W = 0)$ and for $w > 0$,

$$\Pr(W_\Delta \leq -w) = \frac{2\Pr(W \geq w)}{1+\Lambda},\ \Pr(W_\Delta \geq w) = \frac{2\Delta\Pr(W \geq w)}{1+\Lambda},\ \text{for } w>0. \tag{6}$$

The support of W_Δ and W are the same for all $\Delta > 0$. For instance, in §1, the support or set of possible MO-minus-GO differences in weeks with toxicity in year one was the set of integers from −52 to 52, and a family of the form (6) would retain this support as Δ varies. Also, $\Pr(W_\Delta \leq -w) + \Pr(W_\Delta \geq w) = \Pr(W \leq -w) + \Pr(W \geq w) = 2\Pr(W \geq w)$ for each $w > 0$. The family is stochastically ordered by Δ in the sense that $\Delta < \Delta'$ implies $\Pr(W_\Delta \geq w) \leq \Pr(W_{\Delta'} \geq w)$ for all w, so a larger Δ is associated with higher values of W_Δ.

Now (6) implies

$$\Pr(W_\Delta > w) = \Delta\Pr(W_\Delta < -w) \text{ for } w>0. \tag{7}$$

Wolfe (1974) introduced condition (7) calling it 'population weighted symmetry'. In particular, he showed that (7) is a necessary and sufficient condition for the sign of a random variable to be independent of its absolute value, $\text{sign}(W_\Delta) \perp\!\!\!\perp |W_\Delta|$ in Dawid's (1979) notation.

3.3 Model and Amplification

Write $\mathscr{C} = \{(\mathbf{x}_{ij}, u_{ij}), i = 1,\ldots, I, j = 1, 2\}$ for the data on covariates, where again the pairs are matched for $\mathbf{x}$ but not for u. Unlike the model in §2.3, the amplified model conditions on covariates $\mathscr{C}$ rather than on $\mathscr{F}$, viewing both treatment assignment Z_{ij} and response (r_{Tij}, r_{Cij})

as random variables, and it assumes that the dependence between r_{Cij} and Z_{ij} within pair i is due to the failure to match on u_{ij}.

The model has the following components. First, the I distinct pairs, $i = 1,\ldots, I$, are mutually independent of one another given $\mathscr{C}$ for instance, (Y_{Ci}, Z_{i1}, Z_{i2}) and (Y_{Cj}, Z_{j1}, Z_{j2}) are conditionally independent when $i \neq j$. Second, $Y_{Ci} = r_{Ci1} - r_{Ci2}$ and (Z_{i1}, Z_{i2}) are conditionally independent given $\mathscr{C}$ this expresses the idea that any dependence between Y_{Ci} and (Z_{i1}, Z_{i2}) remaining after matching on $\mathbf{x}$ is due to the failure to match also on the unobserved u. The third component is a model relating (Z_{i1}, Z_{i2}) and (u_{i1}, u_{i2}) that is similar to (2), except that it conditions on $\mathscr{C}$ rather than on $\mathscr{F}$ and has parameter λ rather than γ, namely:

$$\Pr(\mathbf{Z}=\mathbf{z} \mid \mathscr{C}) = \prod_{i=1}^{I} \frac{\exp\{\lambda(z_{i1}u_{i1} + z_{i2}u_{i2})\}}{\exp(\lambda u_{i1}) + \exp(\lambda u_{i2})}, \mathbf{u} \in \mathscr{U}. \tag{8}$$

The final component relates Y_{Ci} and (u_{i1}, u_{i2}) using the semiparametric family in §3.2:

$$\Pr(Y_{Ci} \geq y \mid \mathscr{C}) = \exp\{\delta(u_{i1} - u_{i2})\} \Pr(Y_{Ci} \leq -y \mid \mathscr{C}) \text{ for } y > 0. \tag{9}$$

In (9), the distribution of each of the I random variables, Y_{Ci}, $i = 1,\ldots, I$, is a different parametric deformation of possibly different underlying symmetric distributions. Specifically, the degree $\exp\{\delta(u_{i1} - u_{i2})\}$ of asymmetry in Y_{Ci} is determined by the magnitude of difference in the unobserved $u_{i1} - u_{i2}$; it may vary with i. The symmetric distribution that was deformed to produce (9) may also vary with i, so that $\Pr(Y_{Ci} \geq y \mid \mathscr{C}) + \Pr(Y_{Ci} \leq -y \mid \mathscr{C})$ may vary with i. Because (9) describes a relationship among unobservable quantities, it is not a model that might be checked against observed data, but rather a description of the types of statements about unobservable quantities that are warranted by specific results in a sensitivity analysis. Write $\Delta = \exp(\delta)$ and $\Lambda = \exp(\lambda)$.

An amplification is an aid to interpretation, and so it is important to consider the meanings of the parameters Γ, Λ and Δ. The parameters Γ and Λ are similar in form and both refer to the ability to guess the treatment assignments Z_{ij} from certain information; specifically, both refer to bounds on the odds (12) of a bet involving the treatment assignments for two individuals with the same observed covariates, $\mathbf{x}$; however, Γ refers to a guess based on the greater information in $\mathscr{F}$, which includes u_{ij} and r_{Cij} and implicitly $Y_{Ci} = r_{Ci1} - r_{Ci2}$, while Λ refers to a guess based on reduced information, $\mathscr{C}$ which includes u_{ij} but does not include r_{Cij}. It will turn out, perhaps unsurprisingly, that a stronger association Λ based on reduced information corresponds with a weaker association Γ based on more information, even though the forms of the two parameters are similar. Expressed differently, if you were trying to bias the association between treatment assignment and Y_{Ci}, you could create more bias for a given strength of association if you knew Y_{Ci} than if you knew only a variable u_{ij} with which Y_{Ci} is associated. Because $\mathbf{u} \in \mathscr{U}$ the parameter $\exp(|\delta|) = \max(\Delta, 1/\Delta)$ is an upper bound and $\exp(-|\delta|) = \min(\Delta, 1/\Delta)$ is a lower bound on the quantity $\exp\{\delta(u_{i1} - u_{i2})\}$, which is the degree of asymmetry (9) of the difference in outcomes Y_{Ci} produced by the imbalance $u_{i1} - u_{i2}$ in the unobserved covariate. The form of Wolfe's family permits $\exp(|\delta|)$ to be characterized in English as the maximum odds of a positive difference in outcomes under control, $Y_{Ci} > 0$, due to failure to match for the unobserved u. For instance, if $\Delta = 2$, then in a pair matched for $\mathbf{x}$, the individual with the higher u_{ij} might be twice as likely as the individual with the lower u_{ij} to have the higher potential response to control, r_{Cij}. In Wolfe's family, the odds $\Pr(Y_{Ci} > y \mid \mathscr{C}) / \Pr(Y_{Ci} < y \mid \mathscr{C})$ are the same for all y, so within this family,

the parameter Δ may characterized by the value of the odds for any one y, for instance, $y = 0$, that is, $\Pr(Y_{Ci} > 0 \mid \mathscr{C}) / \Pr(Y_{Ci} < 0 \mid \mathscr{C})$.

In thinking about expressions (10) and (11) in Proposition 1 below, recall that when $\Gamma = 1$, $\Pr(\overline{T}_1 \geq k) = \Pr(\overline{\overline{T}}_1 \geq k)$ are both equal to the randomization distribution (1) under the null hypothesis H_0.

Proposition 1 *Under model* (8)–(9) *with fixed* λ *and* δ, *if the null hypothesis* H_0 *of no treatment effect is true, then for a statistic* $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} q_i \operatorname{sign}(V_i Y_i)$, *if* $\delta\lambda \geq 0$

$$\Pr(\overline{T}_1 \geq k) \leq \Pr\{t(\mathbf{Z}, \mathbf{R}) \geq k \mid \mathscr{C}, \mathbf{A}\} \leq \Pr\left(\overline{\overline{T}}_\Gamma \geq k\right) \text{ for all } \mathbf{u} \in \mathscr{U} \tag{10}$$

whereas if $\delta\lambda \leq 0$

$$\Pr(\overline{T}_{1/\Gamma} \geq k) \leq \Pr\{t(\mathbf{Z}, \mathbf{R}) \geq k \mid \mathscr{C}, \mathbf{A}\} \leq \Pr\left(\overline{\overline{T}}_1 \geq k\right) \text{ for all } \mathbf{u} \in \mathscr{U} \tag{11}$$

where

$$\Gamma = \frac{\Delta\Lambda + 1}{\Lambda + \Lambda}. \tag{12}$$

Moreover, the bounds in (10) *and* (11) *are sharp in the sense that they are each attained for particular* $\mathbf{u} \in \mathscr{U}$

Proof. Under the null hypothesis of no effect, H_0, $R_{ij} = r_{Cij}$ and $Y_i = Y_{Ci}$, and these equalities are assumed throughout the proof. Write $\mathbf{Y}_C = (Y_{C1}, \ldots, Y_{CI})^T$, $\mathbf{A}_C = (|Y_{C1}|, \ldots, |Y_{CI}|)^T$, and $\mathbf{S} = \{\operatorname{sign}(Y_{C1}), \ldots, \operatorname{sign}(Y_{CI})\}^T$, and $q_{Ci} = q_i(\mathbf{A}_C) \geq 0$. The model assumed $\mathbf{Y}_C \perp\!\!\!\perp \mathbf{Z} \mid \mathscr{C}$, Wolfe's (1974) Theorem 2.1 applied to Y_{Ci} in (9) implies $\mathbf{S} \perp\!\!\!\perp \mathbf{A}_C \mid \mathscr{C}$, and combining these two facts yields $\mathbf{S} \perp\!\!\!\perp \mathbf{Z} \mid (\mathscr{C}, \mathbf{A}_C)$ by Dawid's (1979) lemma 4. Write $\rho_i = \Pr(V_i = 1 \mid \mathscr{C}) = \Pr(V_i = 1 \mid C, \mathbf{A}_C)$, so that using (8),

$$\rho_i = \frac{\exp\{\lambda(u_{i1} - u_{i2})\}}{1 + \exp\{\lambda(u_{i1} - u_{i2})\}}.$$

Also, write $\eta_{si} = \Pr\{\operatorname{sign}(Y_{Ci}) = s \mid \mathscr{C}\ \mathbf{A}_C\}$ for $s = 1, 0, -1$, so that using (9),

$$\eta_{1i} = \frac{\exp\{\delta(u_{i1} - u_{i2})\}}{1 + \exp\{\delta(u_{i1} - u_{i2})\}} \text{ and } \eta_{-1,i} = \frac{1}{1 + \exp\{\delta(u_{i1} - u_{i2})\}} \text{ if } |Y_{Ci}| > 0, \tag{13}$$

with $\eta_{0i} = 1$ if $|Y_{Ci}| = 0$. If $|Y_{Ci}| > 0$ then $\operatorname{sign}(V_i\ Y_{Ci})$ is 1 or -1 and

$$\Pr\left\{\operatorname{sign}(V_i Y_{Ci}) = 1 \,\middle|\, \mathscr{C}, \mathbf{A}_C\right\} = \eta_{1i}\rho_i + \eta_{-1,i}(1 - \rho_i) = \frac{\exp\{(\delta + \lambda)(u_{i1} - u_{i2})\} + 1}{[1 + \exp\{\delta(u_{i1} - u_{i2})\}][1 + \exp\{\lambda(u_{i1} - u_{i2})\}]};$$

otherwise, if $|Y_{Ci}| = 0$ then $\operatorname{sign}(V_i\ Y_{Ci}) = 0$. Straightforward manipulations then show that if $\delta\lambda \geq 0$ then

$$\frac{1}{2} \leq \Pr\left\{\text{sign}\,(V_i Y_{Ci})=1 \,\middle|\, \mathscr{C}, \mathbf{A}_C\right\} \leq \frac{\Delta\Lambda+1}{(1+\Delta)\,(1+\Lambda)},$$

where the lower bound is attained for $u_{i1} = u_{i2}$ and the upper bound is attained for $u_{i1} = 0$, $u_{i2} = 1$; whereas if $\delta\lambda \leq 0$ then

$$\frac{\Delta\Lambda+1}{(1+\Delta)\,(1+\Lambda)} \leq \Pr\left\{\text{sign}\,(V_i Y_{Ci})=1 \,\middle|\, \mathscr{C}, \mathbf{A}_C\right\} \leq \frac{1}{2},$$

where the lower bound is attained at $u_{i1} = 0$, $u_{i2} = 1$ and the upper bound is attained at $u_{i1} = u_{i2}$. It follows that $\Pr\left\{\sum_{i=1}^{I} q_{Ci}\ \text{sign}\,(V_i Y_i) \geq k \,|\, \mathscr{C}, \mathbf{A}_C\right\}$ is the distribution of the sum of I conditionally independent random variables, where $q_{Ci}\,\text{sign}\,(V_i Y_i) = 0$ if $A_{Ci} = 0$ and otherwise $q_{Ci}\,\text{sign}\,(V_i Y_i) = \pm q_{Ci}$ with probabilities bounded by $\frac{1}{2}$ and $(\Delta\Lambda + 1) / \{(1 + \Delta)\,(1 + \Lambda)\}$. The bounds (10) and (11) for $\sum_{i=1}^{I} q_{Ci}\,\text{sign}\,(V_i Y_i)$ follow immediately from the bounds on the I conditionally independent components by a standard probability inequality (e.g., Ahmed, Leon and Proschan 1981, Lemma 3.3). Finally, for (12),

$$\Gamma=\frac{\Delta\Lambda+1}{\Delta+\Lambda} \iff \frac{\Gamma}{1+\Gamma}=\frac{\Delta\Lambda+1}{(1+\Delta)\,(1+\Lambda)}.$$

The amplification or set of solutions to (12) has the following properties.

Irrelevant covariates: In a familiar way, the value $\Gamma = 1$ for the randomization distribution amplifies into two curves by way of (12), namely $(\Lambda, \Delta) = (1, \Delta)$ for $\Delta \in (0, \infty)$ and $(\Lambda, \Delta) = (\Lambda, 1)$ for $\Lambda \in (0, \infty)$. In words, an unobserved covariate u_{ij} biases the randomization inference only if u_{ij} is relevant to both treatment Z_{ij} and response Y_{Ci} in the absence of treatment.

Limits: Equation (12) defines the correspondence between the one and two parameter sensitivity analyses. It is interesting to note that the one parameter analysis is also the limit of the two parameter analysis, because as $\Delta \to \infty$ in (12), $\Lambda \to \Gamma$. In other words, the one parameter bounds (5) may be understood as a special of the two parameter bounds (10) for a covariate u_{ij} strongly related to Y_{Ci}.

Reduction: If (Λ, Δ) is a solution of (12) then $(\Lambda^{-1}, \Delta^{-1})$ is another solution; this corresponds with replacing u_{ij} by $1 - u_{ij}$. In practice, it will rarely be necessary to discuss both solutions, as they are mirror images.

To know the sign of $\lambda\delta$ in Proposition 1 is to know the direction of the bias induced by u, and this is often not known. If the sensitivity analysis specifies only the magnitudes, $|\lambda|$ and $|\delta|$, and not the signs, then Corollary 2 follows by combining (10) and (11).

Corollary 2 *Under model* (8)–(9) *with fixed* $|\lambda|$ *and* $|\delta|$, *if the null hypothesis* H_0 *of no treatment effect is true, then a statistic* $t(\mathbf{Z}, \mathbf{R}) = \sum_{i=1}^{I} q_i\,\text{sign}\,(V_i Y_i)$ *satisfies the bounds* (5) *with*

$$\Gamma=\frac{\exp(|\lambda|+|\delta|)+1}{\exp(|\lambda|)+\exp(|\delta|)}.$$

Moreover, the bounds are sharp in the sense of being attained for some $\mathbf{u}\in\mathcal{U}$ *and for some signed* λ *and* δ *with the fixed magnitudes* $|\lambda|$ *and* $|\delta|$.

Proposition 1 concerned the hypothesis of no effect, $H_0: r_{Tij}-r_{Cij}=0$ for all ij. To test $H_{\tau_0}: r_{Tij}-r_{Cij}=\tau_{0ij}, \forall ij$, where $\boldsymbol{\tau}_0=(\tau_{011},\ldots,\tau_{0IJ})^T$ is specified, apply Proposition 1 to the adjusted responses, $r_{Cij}=R_{ij}-Z_{ij}-\tau_{0ij}$. This test may be inverted to obtain confidence statements and point estimates (Rosenbaum 1993), for instance for an constant effect, $\boldsymbol{\tau}_0=\theta_0(1,\ldots,1)^T$, or for an attributable effect summarizing nonconstant effects (Rosenbaum 2003, 2007a).

4 Intensity and Toxicity in Ovarian Cancer

For illustration and in light of the discussion in §1, the sensitivity analysis for the 344 matched pair differences toxicity is conducted using three test statistics, $t(\mathbf{Z},\mathbf{R})$. In particular, this illustrates the point that the considerations in §3 apply without change to a wide variety of statistics. The first statistic is Wilcoxon's signed rank statistic, which is familiar and is a good choice for detecting a shift in location.

As noted in the discussion of Figure 1 in §1, it may not be appropriate to describe the differences in toxicity as a shift in location that affects all matched pairs by the same constant amount; rather, the alternative suggested by Conover and Salsburg (1988) may be more appropriate. Specifically, Conover and Salsburg (1988) considered a treatment that has no effect on many people but does affect some people. Their locally most powerful ranks have substantially higher power than Wilcoxon ranks against this alternative. Conover and Salsburg (1988)'s ranks have a form that is not easy to interpret, but they are highly correlated with and practically equivalent to a second set of ranks proposed by Stephenson (1981) who was motivated by different considerations; see Rosenbaum (2007a) for discussion of the relationship. Conceptually (though not computationally), Stephenson suggested looking at all $\binom{I}{m}$ subsets of m of the I pairs of patients, or about 3.9×10^{10} subsets for $I=344$ and $m=5$, or about 5.6×10^{18} subsets for $I=344$ and $m=10$. In effect, Conover and Salsburg suggested using $m=5$. In each subset of m pairs, find the one pair with the largest absolute difference in toxicity. For that one pair, which is the largest of m pairs, a 1 is scored if the higher toxicity was for the MO patient, a positive difference, and a 0 is scored if the higher toxicity was for the GO patient, a negative difference. When Wilcoxon's signed rank statistic is expressed approximately as one of Hoeffding's U-statistics, it is equal to Stephenson's statistic with $m=2$. Taking $m=5$ rather than $m=2$ is saying, in effect, that our interest is in differences in toxicity that are somewhat large, that is, larger than $m-1=4$ others, although in the spirit of U-statistics, the computation is repeated for all subsets of $m=5$. In point of fact, the computations do not involve the $\binom{I}{m}$ subsets directly, and are equivalent to using a signed rank statistic $t(\mathbf{Z},\mathbf{R})=\sum_{i=1}^{I}q_i\,\text{sign}(V_iY_i)$, where if $|Y_i|$ has rank ℓ then $q_i=\binom{\ell-1}{m-1}$, with $\binom{a}{b}$ is defined to equal 0 if $a<b$. That is, if $|Y_i|$ has rank ℓ, then it has the largest absolute difference in $\binom{\ell-1}{m-1}$ of the $\binom{I}{m}$

subsets of m pairs, and $\sum_{i=1}^{I} q_i \text{ sign}(V_i Y_i)$ is linearly related to Stephenson's statistic. In the example, the second and third statistics will be Stephenson's statistic with $m = 5$ or $m = 10$. (Actually, we use the U-statistic form of Wilcoxon's procedure, with $m = 2$, the difference being trivial for $I = 344$.)

Table 1 presents a one-parameter sensitivity analysis in terms of Γ, specifically the upper bound on the one sided significance level from (5) for testing the null hypothesis H_0 of no treatment effect. We describe the one-parameter analysis, then describe its amplification. All three statistics are highly significant when compared to their usual randomization distributions (1) which are equivalent to (5) when Γ = 1. An unobserved covariate strongly related to toxicity that increased the odds of treatment by an MO by 50% or Γ = 1.5 could increase the significance level for Wilcoxon's signed rank statistic to at most 0.036; however, if it doubled the odds of treatment by an MO, Γ = 2, then it could explain the observed association as measured by the Wilcoxon statistic. Because in Figure 1 it appeared that only a small fraction of MO patients experienced substantially increased toxicity, it is not surprising that the results are much less sensitive to bias with $m = 5$ or $m = 10$; that is, relatively high toxicity is more common among MO patients. In particular, when thinking about the largest difference in toxicity in ten pairs, an unobserved covariate strongly related to toxicity would need to more than triple the odds of treatment by an MO to explain the observed association between toxicity and provider specialty.

The analysis in Table 1 is implicitly about an unobserved covariate difference u_{ij} strongly associated with the difference Y_i in toxicity. The amplification reexpresses the very same analysis in terms of two parameters, Δ controlling the association between the covariate u_{ij} and toxicity Y_i, and Λ controlling the association between the covariate u_{ij} and provider specialty $V_i = Z_{i1} - Z_{i2}$. For instance, if Δ = Λ = 2, then u_{ij} can double the odds of treatment from a MO and double the chance of a positive difference in toxicity, $Y_i > 0$, inducing a spurious association between provider speciality and toxicity when there is no actual effect of provider specialty.

As seen in Proposition 1, the sign of λδ determines the direction of the bias. To produce the higher levels of toxicity found among patients of MO's, the unobserved covariate u needs to be either (i) positively associated with both toxicity and treatment by an MO, so Δ > 1 and Λ > 1, or (ii) negatively associated with both toxicity and treatment by an MO, so Δ < 1 and Λ < 1. However, it suffices to consider Δ > 1 and Λ > 1, as (Λ, Δ) and $(\Lambda^{-1}, \Delta^{-1})$ correspond with the same Γ, and the movement from (Λ, Δ) to $(\Lambda^{-1}, \Delta^{-1})$ is the same as replacing u_{ij} by $1 - u_{ij}$.

For fixed Γ > 1 the amplification is the set

$$\mathscr{A}_\Gamma = \left\{(\Lambda, \Delta): \Gamma = \frac{\Delta\Lambda + 1}{\Delta + \Lambda}, \Delta > 0, \Lambda > 0\right\}.$$

For Γ = 1.5, the amplification $\mathscr{A}_{1.5}$ includes (2, 4), (2.618, 2.618), and (4, 2), among many other values. That is, for either Γ = 1.5 in (5) or (Λ, Δ) = (4, 2) in (10), Wilcoxon's signed rank statistic would have the same upper bound 0.036 on the one sided significance level. Here, (Λ, Δ) = (4, 2) refers to a u that quadrupled the odds of treatment by an MO rather than a GO and doubled the odds of greater toxicity. The upper bound on the P-value from Wilcoxon's signed rank statistic is 0.0497 for Γ = 1.532, and Figure 2 depicts the corresponding amplification.

Similarly, for $\Gamma = 2$, the amplification $\mathscr{A}_2$ includes $(\Lambda, \Delta) = (3, 5)$, (3.732, 3.732) and (5, 3). For Stephenson's test with $m = 5$, for a covariate u that was associated with a five-fold increase in the odds of treatment by an MO rather than a GO and a three-fold increase in the odds of greater toxicity, the upper bound on the one-sided significance level is 0.011, the same as for $\Gamma = 2$ in Table 1. For $\Gamma = 3$, the amplification $\mathscr{A}_3$ includes $(\Lambda, \Delta) = (4, 11)$, (5, 7), (5.828, 5.828), (7, 5) and (11, 4), and all of these yield the upper bound of 0.031 in Table 1 for Stephenson's test with $m = 10$. For instance, 0.031 is an upper bound on the significance level if there were an unobserved covariate associated with an eleven-fold increase in the odds of treatment by an MO rather than a GO and associated with a four-fold increase in the odds of greater toxicity.

5 Discussion

When an observational study is discussed in a scientific meeting or journal, it is quite common that a skeptic or critic will raise the possibility that the difference in outcomes is due to failure to control for a specific unobserved covariate u which the skeptic claims is associated with both treatment assignment and outcome. As Bross (1961) emphasized, the skeptic has the responsibility for making this claim specific and plausible in terms of what might reasonably be true of the variable u under discussion. Now it may happen that this unobserved u is very strongly related to the outcome, so the one parameter sensitivity analysis in terms of Γ speaks directly to the issues raised by the skeptic. It may happen, however, that the critic is discussing a covariate u with only a moderate relationship with the outcome, and in this case it is useful to have an amplification that translates the sensitivity analysis in terms of Γ into all corresponding sensitivity analyses in terms of (Λ, Δ). Here, the parameter Λ describes the relationship between an unobserved covariate u and treatment assignment Z, while the parameter Δ describes the relationship between u and the responses r_C that would be exhibited under control. It is convenient that the two-parameter analysis may be determined by simple arithmetic from a reported one-parameter analysis without doing further analysis of the data.

An analysis involving a single sensitivity parameter, here Γ, may be concisely displayed and easily examined, and this is important in empirical papers where many aspects of an investigation must be described in limited space. However, in some contexts, the use of several sensitivity parameters may aid interpretation. An amplification maps a one-dimensional sensitivity analysis into sets of two-dimensional analyses in such a way that the one dimensional analysis accurately describes the corresponding set of two-dimensional analyses; that is, it maps values of Γ into sets $\mathscr{A}_\Gamma$ of values of (Λ, Δ) such that the one analysis for Γ describes all of the analyses for $(\Lambda, \Delta) \in \mathscr{A}_\Gamma$. For instance, if an empirical paper said simply that the upper bound on the significance level for Wilcoxon's signed rank statistic exceeds 0.05 for $\Gamma > 1.532$, then from this one fact a reader could construct Figure 2. In this way, one may calculate and display a one-dimensional analysis, yet have available the interpretations for a two dimensional analysis.

Acknowledgments

This work was supported by grant R01-CA095664 from the U.S. National Cancer Institute and by a grant from the Measurement, Methodology and Statistics Program of the U.S. National Science Foundation.

References

Aakvik A. Bounding a matching estimator: the case of a Norwegian training program. Oxford Bulletin of Economics and Statistics. 2001; 63:115–143.

Ahmed AN, Leon R, Proschan F. Generalized association, with applications in multivariate statistics. Annals of Statistics. 1981; 9:168–176.

Bross IDJ. Statistical criticism. Cancer. 1961; 13:394–400. [PubMed: 13804816]

Brown BM. Symmetric quantile averages and related estimators. Biometrika. 1981; 68:235–242.

Conover WJ, Salsburg DS. Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to 'respond' to treatment. Biometrics. 1988; 44:189–196. [PubMed: 3358987]

Copas J, Eguchi S. Local sensitivity approximations for selectivity bias. Journal of the Royal Statistical Society B. 2001; 63:871–896.

Cornfield J, Haenszel W, Hammond E, Lilienfeld A, Shimkin M, Wynder E. Smoking and lung cancer. Journal of the National Cancer Institute. 1959; 22:173–203. [PubMed: 13621204]

Dawid AP. Conditional independence in statistical theory. Journal of the Royal Statistical Society B. 1979; 41:1–31.

Diprete TA, Gangl M. Assessing bias in the estimation of causal effects. Sociological Methodology. 2004; 34:271–310.

Fisher, RA. The Design of Experiments. Edinburgh: Oliver & Boyd; 1935.

Imbens GW. Sensitivity to exogeneity assumptions in program evaluation. American Economic Review. 2003; 93:126–132.

Gastwirth JL. Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. Jurimetrics. 1992; 33:19–34.

Gastwirth JL, Krieger AM, Rosenbaum PR. Dual and simultaneous sensitivity analysis for matched pairs. Biometrika. 1998; 85:907–920.

McCullagh P. Some applications of quasisymmetry. Biometrika. 1982; 69:303–308.

Maritz J. Exact robust confidence intervals for location. Biometrika. 1979; 66:163–166.

Neyman J. On the application of probability theory to agricultural experiments: Essay on principles, Section 9. Statistical Science. 1923; 5:463–480. reprinted in.

Noether G. Some distribution-free confidence intervals for the center of a symmetric distribution. Journal of the American Statistical Association. 1973; 68:716–719.

Pagano M, Tritchler D. On obtaining permutation distributions in polynomial time. Journal of the American Statistical Association. 1983; 78:435–440.

Robins, JM.; Rotnitzky, A.; Scharfstein, D. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference. In: Halloran, E.; Berry, D., editors. Statistical Models in Epidemiology. NY: Springer; 1999. p. 1-94.

Rosenbaum PR. Sensitivity analysis for certain permutation inferences in matched observational studies. Biometrika. 1987; 74:13–26.

Rosenbaum PR. Hodges-Lehmann point estimates of treatment effect in observational studies. Journal of the American Statistical Association. 1993; 88:1250–1253.

Rosenbaum, PR. Observational Studies. 2nd edition. New York: Springer; 2002.

Rosenbaum PR. Exact confidence intervals for nonconstant effects by inverting the signed rank rest. American Statistician. 2003; 57:132–138.

Rosenbaum PR. Confidence intervals for uncommon but dramatic responses to treatment. Biometrics. 2007a; 63:1164, 1171. [PubMed: 17425641]

Rosenbaum PR. Sensitivity analysis for m-estimates, tests, and confidence intervals in matched observational studies. Biometrics. 2007b; 63:456–464. [PubMed: 17688498]

Rosenbaum PR, Ross RN, Silber JH. Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. Journal of the American Statistical Association. 2007; 102:75–83.

Rosenbaum P, Rubin D. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. Journal of the Royal Statistical Society B. 1983; 45:212–218.

Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology. 1974; 66:688–701.

Silber JH, Rosenbaum PR, Trudeau ME, Chen W, Zhang X, Lorch S, Rapaport-Kelz R, Mosher RE, Even-Shoshan O. Preoperative antibiotics and mortality in the elderly. Annals of Surgery. 2005; 242:107–114. [PubMed: 15973108]

Silber JH, Rosenbaum PR, Polsky D, Ross RN, Even-Shoshan O, Schwartz S, Armstrong KA, Randall TC. Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? Journal of Clinical Oncology. 2007; 25:1169–1175. [PubMed: 17401005]

Slade EP, Stuart EA, Alkever DSS, Karakus M, Green KM, Ialongo N. Impacts of age of onset of substance use disorders on risk of adult incarceration among disadvantaged urban youth. Drug and Alcohol Dependence. 2008; 95:1–13. [PubMed: 18242006]

Stephenson WR. A general class of one-sample nonparametric test statistics based on subsamples. Journal of the American Statistical Association. 1981; 76:960–966.

Wang L, Krieger AM. Causal conclusions are most sensitive to unobserved binary covariates. Statistics in Medicine. 2006; 25:2257–2271. [PubMed: 16220480]

Welch BL. On the z-test in randomized blocks. Biometrika. 1937; 29:21–52.

Wolfe DA. A characterization of population weighted symmetry and related results. Journal of the American Statistical Association. 1974; 69:819–822.

Yanagawa T. Case-control studies: assessing the effect of a confounding factor. Biometrika. 1984; 71:191–194.

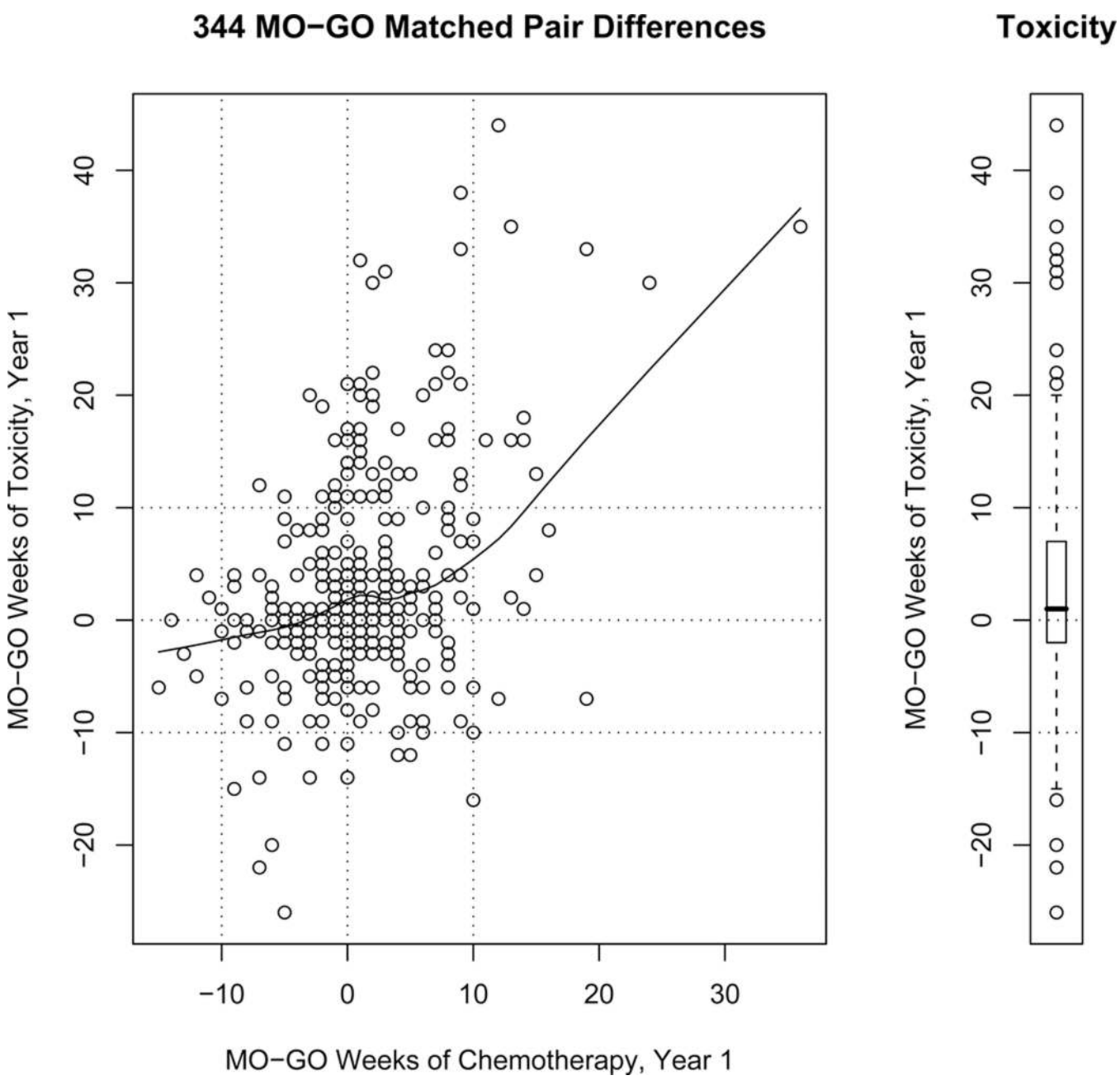

Figure 1.
Matched Pair Difference, MO-GO, in Toxicity Weeks Plotted Against Matched Pair Difference in Chemotherapy Weeks in the Year After Diagnosis for 344 Pairs of Patients. The curve is a lowess smooth, and the boxplot displays the marginal distribution of the MO-GO difference in toxicity.

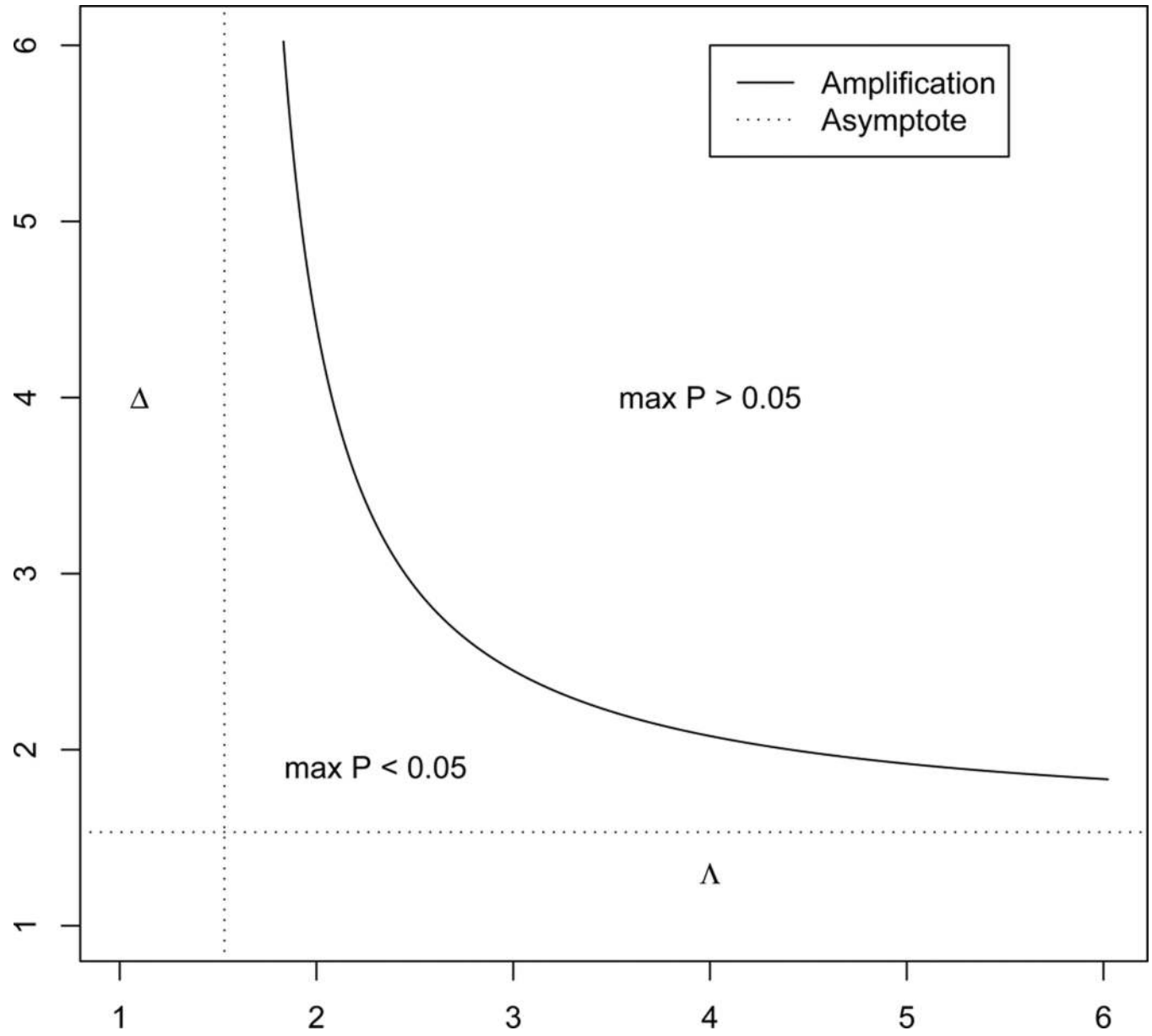

Figure 2.
Amplification (Λ, Δ) of Γ = 1.532. In the ovarian data, Wilcoxon's signed rank statistic has a maximum P-value of 0.05 at Γ = 1.532. The dotted lines are the twin asymptotes of Λ = 1.532 and Δ = 1.532.

Table 1

Upper Bounds on One-Sided Significance Levels Testing for No Effect On Toxicity in Year 1.

Γ	Wilcoxon $m = 2$	Stephenson $m = 5$	Stephenson $m = 10$
1	6.3×10^{-7}	5.0×10^{-9}	1.8×10^{-8}
1.5	0.036	0.00013	0.000027
2	0.62	0.011	0.0010
2.5		0.11	0.0080
3			0.031