Split Samples and Design Sensitivity in Observational Studies

Ruth Heller, Paul R. Rosenbaum and Dylan S. Small

University of Pennsylvania, Philadelphia

Abstract. An observational or nonrandomized study of treatment effects may be biased by failure to control for some relevant covariate that was not measured. The design of an observational study is known to strongly affect its sensitivity to biases from covariates that were not observed. For instance, the choice of an outcome to study, or the decision to combine several outcomes in a test for coherence can materially affect the sensitivity to unobserved biases. Decisions that shape the design are, therefore, critically important, but they are also difficult decisions to make in the absence of data. We consider the possibility of randomly splitting the data from an observational study into a smaller planning sample and a larger analysis sample, where the planning sample is used to guide decisions about design. After reviewing the concept of design sensitivity, we evaluate sample splitting in theory, by numerical computation, and by simulation, comparing it to several methods that use all of the data. Sample splitting is remarkably effective, much more so in observational studies than in randomized experiments: splitting 1000 matched pairs into 100 planning pairs and 900 analysis pairs often materially improves the design sensitivity. An example from genetic toxicology is used to illustrate the method.

Keywords: Coherence; multiple comparisons; permutation test; sensitivity analysis.

1 Introduction: Design to Reduce Sensitivity to Unobserved Bias

The design of an observational study entails choosing the circumstances in which the study will be conducted and developing a protocol for its analysis (Rosenbaum 1999).

Many features of the design of an observational study affect its sensitivity to covariates that were not measured, including the pattern and influence of doses, the way doses are incorporated in the analysis, the use of coherent multivariate responses, the heterogeneity of experimental material, and the strength of instrumental variables; see Rosenbaum (2004, 2005) and Small and Rosenbaum (2008). Alas, these features are often of uncertain form before data are obtained, so an analysis plan that seeks improved design by making guesses about these features may guess incorrectly, yielding an inferior design. This raises the possibility of splitting the sample, using the first portion to plan the analysis based on the second portion.

In considering questions of this type, a useful tool is the design sensitivity (Rosenbaum 2004). The *design sensitivity* is a number, $\widetilde{\Gamma}$, that evaluates the design of an observational study, that is, a particular data generating process and planned protocol for analysis. Once data have been collected in an observational study, a sensitivity analysis asks: How far would this study have to depart from a randomized experiment to alter the qualitative conclusions? The design sensitivity, $\widetilde{\Gamma}$, anticipates the outcome of a sensitivity analysis, in much the same way that the power of a test anticipates the outcome of the test. A better design has a larger design sensitivity, $\widetilde{\Gamma}$; it is expected to be less sensitive to unobserved biases if the treatment is effective and biases are absent. The design sensitivity is a basis for appraising competing designs for observational studies. The current paper considers the possibility of using a split sample to make choices that increase the design sensitivity.

Cox (1975) used split samples to select one of several hypotheses to test in a context, such as a randomized trial, in which efficiency rather than bias is the central

concern. He found that split samples ran a close second in terms of power to multiple comparisons based on the Bonferroni inequality, but he observed that split samples were more flexible, and perhaps more easily adapted to complex settings. When thinking about sensitivity to unobserved biases, however, split samples outperform multiple comparisons based on the Bonferroni inequality, as is seen in §3.

As an example, consider the following study in genetic toxicology. Masjedi, et al. (2000) examined genetic damage from tuberculosis and the anti-tuberculosis drugs used to treat it using $I = 36$ pairs of a patient and a healthy control matched for age and gender. All were nonsmokers. As is common in genetic toxicology, they evaluated genetic damage in lymphocyte cultures, using two measures, the number of chromosome aberrations (CA), excluding gaps, per 100 cells, and the frequency of micronuclei (MN) per 1000 cells. We focus on their comparison of patients to controls. (They also compare tuberculosis patients before and after drug treatment, to separate the effects of drugs from the effects of tuberculosis.)

Although subjects were matched for age, gender and smoking, there is little to ensure that treated patients and untreated matched controls were similar in other ways. If the treated patients and matched controls differed in terms of a relevant unmeasured covariate, then they might have differing outcomes, differing levels of CA and MN, for reasons unrelated to tuberculosis and the anti-tuberculosis drugs used to treat it. A sensitivity analysis asks what such an unobserved covariate would have to be like to alter the qualitative conclusions of the study; see §2.2 for a review of sensitivity analysis. In this rather simple example, one has three choices, namely making CA or MN the primary outcome, or seeking a coherent multivariate

pattern of associations of CA and MN jointly with treatment; see §4 for discussion of coherence. Although this decision will affect the study's ultimate sensitivity to bias from unmeasured covariates, it is not an easy decision to make in the absence of data. One cannot perform a large number of analyses and report only the most promising analyses; that strategy would give misleading conclusions even in a randomized trial. In §5, we split the sample at random into a planning sample of 6 pairs and an analysis sample of 30 pairs. The planning sample of 6 pairs guides the decision among the three choices; then, the analysis sample of 30 pairs conducts a sensitivity analysis with that choice. This works well in the small example, and more importantly, the theoretical results in §3 and §4 suggest it ought to work increasingly well with increasing sample size, and hence it is a useful strategy in the larger studies more often encountered in practice. In §6, we discuss splitting for other design decisions.

2 Notation and Review

2.1 Randomized experiments and the effects caused by treatments

There are I pairs, $i = 1, \ldots, I$, of two individuals, $j = 1, 2$, one treated, denoted $Z_{ij} = 1$, the other control, denoted $Z_{ij} = 0$, so $Z_{i1} + Z_{i2} = 1$ for all i. Matching has controlled an observed pretreatment covariate $\mathbf{x}_{ij}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$, but may fail to control an unobserved covariate, u_{ij}, so typically $u_{i1} \neq u_{i2}$. The jth individual in pair i has two potential K-dimensional vector responses, $\mathbf{r}_{Tij} = (r_{Tij1}, \ldots, r_{TijK})^T$ and $\mathbf{r}_{Cij} = (r_{Cij1}, \ldots, r_{CijK})^T$, where $\mathbf{r}_{Tij}$ is observed if this individual receives treatment, $Z_{ij} = 1$, or $\mathbf{r}_{Cij}$ is observed if this individual receives control, $Z_{ij} = 0$, and the

effect of the treatment on this individual, $\mathbf{r}_{Tij} - \mathbf{r}_{Cij}$, cannot be calculated from the observed treatment assignment, Z_{ij}, and observed response $\mathbf{R}_{ij} = Z_{ij}\,\mathbf{r}_{Tij} + (1 - Z_{ij})\,\mathbf{r}_{Cij}$; see Neyman (1923) and Rubin (1974). In §1, $I = 36$, $\mathbf{x}_{ij}$ records age and gender, and there are $K = 2$ outcomes, namely CA and MN. Write $\mathcal{F} = \{(\mathbf{r}_{Tij}, \mathbf{r}_{Cij}, \mathbf{x}_{ij}, u_{ij}),\ i = 1, \ldots, I,\ j = 1, 2\}$, $\mathbf{Z} = (Z_{11}, Z_{12}, \ldots, Z_{I2})^T$ and $\mathcal{Z}$ for the set of possible values of $\mathbf{Z}$; i.e., $\mathbf{z} \in \mathcal{Z}$ if and only if $z_{ij} = 0$ or 1 and $z_{i1} + z_{i2} = 1$. Write $|A|$ for the number of elements in a finite set A, so $|\mathcal{Z}| = 2^I$.

One might focus on a specific scalar aspect of the response, defined by a function $y : \mathbb{R}^K \to \mathbb{R}$ specified in advance of examination of the data. The null hypothesis that this aspect is not affected asserts $H_0^y : y(\mathbf{r}_{Tij}) = y(\mathbf{r}_{Cij}),\ \forall i, j$, or equivalently $H_0^y : y_{Tij} = y_{Cij}\ \forall i, j$, if we write $y_{Tij} = y(\mathbf{r}_{Tij})$ and $y_{Cij} = y(\mathbf{r}_{Cij})$. Also, write $Y_{ij} = Z_{ij}\,y_{Tij} + (1 - Z_{ij})\,y_{Cij}$, and $\mathbf{Y} = (Y_{11}, Y_{12}, \ldots, Y_{I2})^T$. For the moment, $y(\cdot)$ is a function specified in advance of the data, but we will soon be interested in splitting the sample, picking $y(\cdot)$ based on one part of the sample, and testing H_0^y using the other part of the sample. In §1, three choices of $y(\cdot)$ were considered, one that selects CA as the primary outcome, one that selects MN as the primary outcome, and one that combines them into a coherent unidimensional summary.

In a randomized paired experiment, treatment assignments would be determined by I independent flips of a fair coin, so that $\Pr(Z_{i1} = 1 \mid \mathcal{F}) = \frac{1}{2}$, $Z_{i2} = 1 - Z_{i1}$, independently in distinct pairs. (All probabilities implicitly condition on the design requirement, $\mathbf{Z} \in \mathcal{Z}$, but the notation does not indicate this explicitly.) Randomization would form a basis for testing H_0^y. Under H_0^y, $Y_{ij} = y_{Tij} = y_{Cij}$ is a function

of $\mathcal{F}$, and is therefore fixed by conditioning on $\mathcal{F}$, so

$$\Pr\left\{t\left(\mathbf{Z},\mathbf{Y}\right)\geq c\,|\,\mathcal{F}\right\}=\left|\left\{\mathbf{z}\in\mathcal{Z}:t\left(\mathbf{z},\mathbf{Y}\right)\geq c\right\}\right|/2^{I}. \tag{1}$$

This yields the usual null distribution of Wilcoxon's signed rank statistic, $t\left(\mathbf{Z},\mathbf{Y}\right)=\sum_{i=1}^{I}\operatorname{sgn}\left\{\left(Y_{i1}-Y_{i2}\right)\left(Z_{i1}-Z_{i2}\right)\right\}\operatorname{rank}\left(\left|Y_{i1}-Y_{i2}\right|\right)$ where $\operatorname{sgn}\left(w\right)=1$, $\frac{1}{2}$, or 0 as $w>0$, $w=0$, or $w<0$ and $\operatorname{rank}\left(\cdot\right)$ is ranking with average ranks for ties.

2.2 Observational studies and sensitivity to bias from an unobserved covariate

In an observational study, matching may fail to control a relevant unobserved covariate, so $\Pr\left(Z_{i1}=1\,|\,\mathcal{F}\right)$ deviates from $\frac{1}{2}$. A simple model for sensitivity analysis in an observational study asserts that the odds of treatment deviates from 1 by at most a factor of $\Gamma\geq1$,

$$\frac{1}{\Gamma}\leq\Pr\left(Z_{i1}=1\,|\,\mathcal{F}\right)/\Pr\left(Z_{i1}=0\,|\,\mathcal{F}\right)\leq\Gamma,\;\;i=1,\ldots,I, \tag{2}$$

with independent assignments in distinct pairs. It is straightforward to show that the family of models for treatment assignment permitted by (2) is the same as the family of models with an unobserved u_{ij}, $0\leq u_{ij}\leq1$, $\mathbf{u}=\left(u_{11},\ldots,u_{I2}\right)^{T}$, such that

$$\Pr\left(\mathbf{Z}=\mathbf{z}\,|\,\mathcal{F}\right)=\prod_{i=1}^{I}\frac{\exp\left(\gamma\sum_{j=1}^{2}z_{ij}\,u_{ij}\right)}{\exp\left(\gamma\,u_{i1}\right)+\exp\left(\gamma\,u_{i2}\right)}\text{ for }\mathbf{z}\in\mathcal{Z}\text{ with }\mathbf{u}\in\mathcal{U}, \tag{3}$$

where $\gamma=\log\left(\Gamma\right)$ and $\mathcal{U}=\left[0,1\right]^{2I}$ is the $2I$-dimensional unit cube, because (3) straightforwardly implies (2), and if (2) holds with independence between pairs, then

define $u_{i1} = \log\left\{\Pr\left(Z_{i1}=1 \mid \mathcal{F}\right) / \Pr\left(Z_{i1}=0 \mid \mathcal{F}\right)\right\}/\gamma$ and $u_{i2}=0$ so that (3) holds; see Rosenbaum (2002, §4) for details and extensions. For $\Gamma = 1$ or $\gamma = \log(\Gamma) = 0$, (3) yields the randomization distribution, $\Pr(\mathbf{Z}=\mathbf{z}) = 2^{-I}$ for $\mathbf{z} \in \mathcal{Z}$, and the null randomization distribution (1). For fixed $\Gamma > 1$ or $\gamma = \log(\Gamma) > 0$, the distribution of treatment assignments $\Pr\left(\mathbf{Z}=\mathbf{z} \mid \mathcal{F}\right)$ is unknown to bounded degree, from which it is possible to produce bounds on $\Pr\left\{t\left(\mathbf{Z},\mathbf{Y}\right) \geq c \mid \mathcal{F}\right\}$ and on associated inference quantities, such as significance levels, confidence intervals and point estimates. A sensitivity analysis computes these bounds for several values of Γ, thereby indicating the degree to which conclusions might be altered by unobserved biases of various magnitudes. For instance, under H_0^y and (3), for each fixed $\Gamma \geq 1$, a critical value, c_Γ, may be determined such that $\max_{\mathbf{u}\in\mathcal{U}} \Pr\left\{t\left(\mathbf{Z},\mathbf{Y}\right) \geq c_\Gamma \mid \mathcal{F}\right\} = \alpha$, so if $t\left(\mathbf{Z},\mathbf{Y}\right) \geq c_\Gamma$ is observed, then a bias of magnitude Γ would not alter the conclusion that H_0^y is rejected at level α. For Wilcoxon's signed rank statistic without ties, $c_\Gamma \doteq \kappa I\left(I+1\right)/2 + \Phi^{-1}\left(1-\alpha\right)\sqrt{\kappa\left(1-\kappa\right)I\left(I+1\right)\left(2I+1\right)/6}$ for large I where $\kappa = \Gamma/\left(1+\Gamma\right)$ and $\Phi^{-1}\left(\cdot\right)$ is the inverse of the standard Normal cumulative distribution; see Rosenbaum (1987; 2002, §4.3.3; 2005, §2.3). (In general, the critical value c_Γ may depend on $\mathcal{F}$ as well as Γ, but the notation does not indicate this. For many rank statistics, such as Wilcoxon's statistic, c_Γ depends on I and Γ but not on $\mathcal{F}$.)

For other sensitivity analyses, see Cornfield, et al. (1959), Breslow and Day (1980, §2.7), Rosenbaum and Rubin (1983), Gastwirth (1992), Marcus (1997), Lin, et al. (1998), Robins, et al. (1999), Copas and Eguchi (2001), and Imbens (2003). For applications, see Aakvik (2001), Diprete and Gangl (2004), and Silber, et al. (2005).

2.3 Design sensitivity

Some designs for observational studies are more resistant to unobserved biases than others, and it is useful to have a quantitative measure of this. The design and analytic protocol for an observational study may be evaluated in terms either of the power of a sensitivity analysis or a simpler related quantity, the design sensitivity. For fixed Γ, when H_0^y is false and some specific alternative is true instead, the *power of a sensitivity analysis* is the chance that $t(\mathbf{Z}, \mathbf{Y}) \geq c_\Gamma$. In principle, the alternative hypothesis might specify $\mathcal{F}$ in detail, in which case the power is $\Pr\{t(\mathbf{Z}, \mathbf{Y}) \geq c_\Gamma \mid \mathcal{F}\}$, but it is conceptually and practically simpler to describe the alternative in terms of a model that generates $\mathcal{F}$, so the power becomes $E[\Pr\{t(\mathbf{Z}, \mathbf{Y}) \geq c_\Gamma \mid \mathcal{F}\}]$, the expectation being taken over the model that generates $\mathcal{F}$. For $\Gamma = 1$, this reproduces the usual definition of the power of a randomization test.

If the situation were favorable, in the sense that the treatment was effective and there was no bias from unobserved covariates, then we would not know this from the observed data. We would see that treated subjects had higher responses than controls, but we would be uncertain whether this was an effect caused by the treatment or bias from some unobserved covariate. The best we could hope to report is the ostensible effect of the treatment is insensitive to small and moderate biases. The chance that this hope is realized is the power of the sensitivity analysis computed assuming this favorable situation. Under the favorable situation, the power of the sensitivity analysis is the chance that $t(\mathbf{Z}, \mathbf{Y}) \geq c_\Gamma$ when $\mathbf{Z}$ is randomized under a conventional model for a treatment effect, such as independent and identically distributed sampling. In this case, the chance that $t(\mathbf{Z}, \mathbf{Y}) \geq c_\Gamma$ may be computed for

Wilcoxon's signed rank statistic using standard power computations (e.g., Lehmann 1975, §4.2) applied to the nonstandard critical value, c_Γ. Specifics follow. Consider the favorable situation with a treatment effect and no bias from unobserved covariates, with the additional assumption that the treated minus control differences $D_i = (2Z_{i1} - 1)(Y_{i1} - Y_{i2})$ are independent and identically distributed. Defining $p = \Pr(D_i > 0)$, $p_1^{'} = \Pr(D_i + D_j > 0)$ and $p_2^{'} = \Pr(D_i + D_j > 0 \wedge D_i + D_k > 0)$ with $i < j < k$, Lehmann (1975, §4.2) shows the nonnull expectation μ_y and variance σ_y^2 of the signed rank statistic $t(\mathbf{Z}, \mathbf{Y})$ are $\mu_y = I(I-1)p_1^{'}/2 + Ip$ and

$$\sigma_y^2 = I(I-1)(I-2)\left(p_2^{'} - p_1^{'\,2}\right) + \frac{I(I-1)}{2}\left\{2\left(p - p_1^{'}\right)^2 + 3p_1^{'}\left(1 - p_1^{'}\right)\right\} + Ip(1-p),$$

so that the central limit theorem yields the approximate power of a one-sided sensitivity analysis as $\Pr\{t(\mathbf{Z}, \mathbf{Y}) \geq c_\Gamma\} \approx 1 - \Phi\left\{\left(c_\Gamma - \mu_y\right)/\sigma_y\right\}$.

In the favorable situation, under mild conditions as $I \rightarrow \infty$ with a fixed treatment effect, the power of the sensitivity analysis tends to 1 for small values of $\Gamma \geq 1$ and to zero for large values of Γ; see Rosenbaum (2004, 2005). The transition from limiting power 1 to limiting power 0 occurs at a value of Γ, say $\widetilde{\Gamma}$, called the *design sensitivity*. This says that once sampling variability has been driven out by letting $I \rightarrow \infty$, a particular study design and treatment effect can be distinguished from all biases $\Gamma < \widetilde{\Gamma}$ but not from biases $\Gamma > \widetilde{\Gamma}$. Other things being equal, one prefers a design with a larger design sensitivity, $\widetilde{\Gamma}$. In the case of Wilcoxon's signed rank statistic, $\widetilde{\Gamma}$ has a simple explicit form, namely $\widetilde{\Gamma} = p_1^{'}/\left(1 - p_1^{'}\right)$; see §3.1 below. The design sensitivity describes the limit as $I \rightarrow \infty$, but like Pitman efficiency, it tends to provide an accurate relative ordering of situations for moderately large I; see Table

4 in Rosenbaum (2004) and Tables 3-6 in Small and Rosenbaum (2008).

3 Split Samples and Design Sensitivity: Selecting an Outcome

3.1 Splitting reduces power, but does not reduce design sensitivity

Consider splitting the sample at random into two parts of size $(1-\zeta)I$ and ζI, $0<\zeta<1$, using the planning sample of size $(1-\zeta)I$ as the basis for an empirical choice of $y(\cdot)$ in §2.1, which is then used as the primary outcome in a sensitivity analysis in the analysis sample of size ζI. For instance, $y(\cdot)$ might focus attention on one outcome or combine several outcomes into a single measure. Typically, $y(\cdot)$ would be chosen with a view to increasing the design sensitivity, say by picking the one outcome that appears most dramatically affected in the planning sample. A fair comparison of the full sample of I pairs and the analysis sample of ζI must take account of the possibility that the planning sample of $(1-\zeta)I$ pairs may yield a better choice of $y(\cdot)$. Later sections consider various fair comparisons in simple, stylized settings. However, in the current section we consider an unfair comparison; specifically, we take $y(\cdot)$ as fixed, and compare using the same $y(\cdot)$ with samples of size I or ζI. As will be seen, for a given $y(\cdot)$, the switch from I to ζI reduces the power of the sensitivity analysis but leaves the design sensitivity unchanged.

As discussed in §2.3, as $I\rightarrow\infty$, the power of the signed rank test applied to $y(\mathbf{R})=Y$ is approximately $\Pr\left\{t(\mathbf{Z},\mathbf{Y})\geq c_{\Gamma}\right\}\approx 1-\Phi\left\{\left(c_{\Gamma}-\mu_{y}\right)/\sigma_{y}\right\}$ where

$\left(c_\Gamma - \mu_y\right)/\sigma_y$ is

$$= \frac{\kappa I\left(I+1\right)/2 + \Phi^{-1}\left(1-\alpha\right)\sqrt{\kappa\left(1-\kappa\right)I\left(I+1\right)\left(2I+1\right)/6} - I\left(I-1\right)p_1^{'}/2 - Ip}{\sqrt{I\left(I-1\right)\left(I-2\right)\left(p_2^{'} - p_1^{'\,2}\right) + \frac{I(I-1)}{2}\left\{2\left(p-p_1^{'}\right)^2 + 3p_1^{'}\left(1-p_1^{'}\right)\right\} + Ip\left(1-p\right)}}$$

$$\approx \frac{\sqrt{I}\left(\kappa - p_1^{'}\right)}{2\sqrt{p_2^{'} - p_1^{'\,2}}} \text{ with } \kappa = \frac{\Gamma}{1+\Gamma}, \tag{4}$$

so $\Pr\left\{t\left(\mathbf{Z},\mathbf{Y}\right) \geq c_\Gamma\right\} \rightarrow 1$ for $\Gamma < \widetilde{\Gamma} = p_1^{'}/\left(1-p_1^{'}\right)$ and $\Pr\left\{t\left(\mathbf{Z},\mathbf{Y}\right) \geq c_\Gamma\right\} \rightarrow 0$ for $\Gamma > \widetilde{\Gamma} = p_1^{'}/\left(1-p_1^{'}\right)$. If the number I of pairs is reduced to ζI, $0 < \zeta < 1$, then $\left|c_\Gamma - \mu_y\right|/\sigma_y$ is reduced in magnitude by approximately a factor of $\sqrt{\zeta}$, so when $\kappa - p_1^{'} < 0$ or equivalently $\Gamma < \widetilde{\Gamma}$, the power $1 - \Phi\left\{\left(c_\Gamma - \mu_y\right)/\sigma_y\right\}$ of the sensitivity analysis is reduced, but the limiting behavior as $I \rightarrow \infty$, and hence the numerical value $\widetilde{\Gamma}$ of the design sensitivity are unchanged. For instance, if the planning sample is $1 - \zeta = 1/10$ of the total, then $\left|c_\Gamma - \mu_y\right|/\sigma_y$ is reduced by approximately a factor$\sqrt{\zeta} = 0.949$, whereas if $1 - \zeta = 1/3$ then $\sqrt{\zeta} = 0.82$. However, as $I \rightarrow \infty$, for both the full sample of I pairs and the analysis sample of ζI pairs, the power tends to 1 for $\Gamma < \widetilde{\Gamma}$ and to 0 for $\Gamma > \widetilde{\Gamma}$.

The planning sample is used to pick a good $y\left(\cdot\right)$ in the hope of increasing the design sensitivity $\widetilde{\Gamma}$ when the analysis is performed on the analysis sample. The key conclusion of the current section is that, as $I \rightarrow \infty$, even a modest increase in the design sensitivity $\widetilde{\Gamma}$ by an improved choice of $y\left(\cdot\right)$ will ultimately dominate use of the entire sample with a slightly inferior choice of $y\left(\cdot\right)$, because the power function, viewed as a function of Γ, is tending to a step function with a single step at $\widetilde{\Gamma}$. As $I \rightarrow \infty$, the location of $\widetilde{\Gamma}$ is all important.

With L choices for the function $y(\cdot)$, an alternative to sample splitting is to use all I pairs and correct for multiple testing using the Bonferroni inequality or a related procedure. In (4), this means replacing α by α/L. From (4), replacing α by α/L affects the power but does not alter the design sensitivity, $\widetilde{\Gamma}$. Although we compare the power of splitting and Bonferroni in §3.3 and §4.3, the two approaches are not completely comparable. In the current context, use of the Bonferroni inequality yields a test of the global null hypothesis that none of the outcomes are affected, H_0, while splitting is selecting one hypothesis to test, H_0^y.

In short, a good choice of $y(\cdot)$ can yield a larger design sensitivity, $\widetilde{\Gamma}$; however, the design sensitivity is the same numerical value for: (i) picking $y(\cdot)$ based on a priori considerations, (ii) picking the same $y(\cdot)$ based on sample splitting, and (iii) picking the same $y(\cdot)$ using all of the data and correcting using the Bonferroni inequality. For finite I, the powers of these three procedures are different. In an experiment, with $\Gamma = 1$, this is analogous to saying that all three procedures yield consistent tests, but their powers are different. In §3.2 and §3.3, the numerical power of a sensitivity analysis with $\Gamma \geq 1$ is determined in simple settings. In §3.2, sample splitting is contrasted with an a priori guess in choosing between two possible $y(\cdot)$'s. In §3.3, sample splitting is contrasted with the Bonferroni inequality.

3.2 Sample splitting versus an a priori choice: two outcomes

In the current section, there are $K = 2$ outcomes, the investigator will choose one primary outcome, and will perform a sensitivity analysis for the primary outcome using the signed rank statistic. Two strategies are contrasted. In the first strategy,

the investigator splits the sample at random in fractions $(1-\zeta)\,I$ and ζI, $0<\zeta<1$, picks the outcome with the larger estimated design sensitivity in the $(1-\zeta)\,I$ fraction, and applies the sensitivity analysis to that outcome in the remaining ζI sample. In the second strategy, without splitting the sample, the investigator guesses which outcome is best, guessing correctly with probability Υ, incorrectly with probability $1-\Upsilon$, and performs the analysis on the full sample of I pairs using the guessed outcome. Without loss of generality, it is assumed that the first outcome has the larger design sensitivity, but of course the investigator does not know this.

Let T_{k1}, T_{k2}, and $T_{k\cdot}$ be, respectively, the signed rank statistics from the $(1-\zeta)\,I$ split, the ζI split, and the full sample of size I, for outcome k. It is easy to verify that the first outcome is estimated to have larger design sensitivity if $T_{11}-T_{21}>0$, and in that case the first strategy uses T_{12}; otherwise, it uses T_{22}. Under very mild conditions, as $I\rightarrow\infty$, the distributions of the six signed rank statistics tend to Normal distributions with expectation and variance given by the expressions parallel to those for μ_y and σ_y^2 in §2.3, but with I replaced appropriately by $(1-\zeta)\,I$ or ζI. The Normal approximation is used in computing the power of the sensitivity analysis. Write $H=1$ if $T_{11}-T_{21}>0$, $H=2$ otherwise, and $\widehat{T}=T_{H,2}$, so $\widehat{T}$ is the signed rank statistic in the analysis sample for the outcome chosen by the planning sample. In parallel, write $\widetilde{T}$ for $T_{1\cdot}$ or $T_{2\cdot}$, picking independently of $(T_{1\cdot},T_{2\cdot})$ the outcome with the larger true design sensitivity, namely outcome 1, with probability Υ, and outcome 2 with probability $1-\Upsilon$. For a fixed Γ and $\kappa=\Gamma/(1+\Gamma)$, the first strategy has approximate power $\Pr\left(\widehat{T}\geq c_{\Gamma,\zeta}\right)$ where $c_{\Gamma,\zeta}\doteq\kappa\zeta I\left(\zeta I+1\right)/2+\Phi^{-1}\left(1-\alpha\right)\sqrt{\kappa\left(1-\kappa\right)\zeta I\left(\zeta I+1\right)\left(2\zeta I+1\right)/6}$, whereas the second

strategy has power $\Pr\left(\widetilde{T} \geq c_\Gamma\right) = \Upsilon \Pr\left(T_{1\cdot} \geq c_\Gamma\right) + (1-\Upsilon)\Pr\left(T_{2\cdot} \geq c_\Gamma\right)$ where $c_\Gamma = c_{\Gamma,1}$ was defined in §2.2.

The limiting case, as $I \to \infty$, is elementary and reinforces the discussion in §3.1. Write $\widetilde{\Gamma}_1$ and $\widetilde{\Gamma}_2$ for the design sensitivity for $T_{1\cdot}$ and $T_{2\cdot}$, respectively. For $1 \leq \Gamma < \min\left(\widetilde{\Gamma}_1, \widetilde{\Gamma}_2\right)$, the power of the sensitivity analysis tends to 1 for all $0 < \zeta < 1$ and $0 < \Upsilon < 1$, whereas for $\Gamma > \max\left(\widetilde{\Gamma}_1, \widetilde{\Gamma}_2\right) \geq 1$, the power of the sensitivity analysis tends to 0 for all $0 < \zeta < 1$ and $0 < \Upsilon < 1$. If $\widetilde{\Gamma}_1 > \Gamma > \widetilde{\Gamma}_2 \geq 1$, then the choice of outcome matters for the limiting power, and as $I \to \infty$, $\Pr\left(\widehat{T} \geq c_{\Gamma,\zeta}\right) \to 1$ whereas $\Pr\left(\widetilde{T} \geq c_\Gamma\right) \to \Upsilon$. This happens because $\widehat{T}$ is very likely to choose the correct outcome for sufficiently large $(1-\zeta)I$, but $\widetilde{T}$ chooses the right outcome with probability Υ, yielding limiting power 1, or the wrong outcome with probability $1-\Upsilon$, yielding limiting power 0. The case $\widetilde{\Gamma}_1 > \Gamma = \widetilde{\Gamma}_2 = 1$, is just slightly different, with $\Pr\left(\widehat{T} \geq c_{\Gamma,\zeta}\right) \to 1$ and $\Pr\left(\widetilde{T} \geq c_\Gamma\right) \to \Upsilon + (1-\Upsilon)\alpha$, because in this case there is still an α chance of rejection when the better outcome is not chosen. Despite winning as $I \to \infty$, splitting affects power for finite I, which we now investigate.

Write $V_{ik} = (R_{i1k} - R_{i2k})(Z_{i1} - Z_{i2})$ for the treated-minus-control difference for outcome k in pair i, and $\mathbf{V}_i = (V_{i1}, \ldots, V_{iK})^T$. To examine power for finite I, the $K = 2$ outcomes will be independent of each other with additive effect $\omega_k \tau_k$, so $r_{Tijk} = r_{Cijk} + \omega_k \tau_k$, with $V_{ik} = \omega_k \tau_k + (r_{Ci1k} - r_{Ci2k})(Z_{i1} - Z_{i2}) \sim_{iid} N(\omega_k \tau_k, \omega_k^2)$, so that τ_k is the magnitude of the effect in units of the standard deviation ω_k of the matched pair difference. By invariance, the power for this model is the same as the power for the special case $\omega_k = 1$, $V_{ik} \sim N(\tau_k, 1)$, $k = 1, 2$. In Table 1, $\tau_1 = \frac{1}{2}$ and $\tau_2 = \frac{1}{4}$ or $\tau_2 = 0$, the sample size is $I = 50$ or 100 or 500 or 1000, and the guesses

Table 1: Power by Splitting or Guessing, $\Upsilon = 2/3$. Values are power for $(\widehat{T}, \widetilde{T})$.

$(\tau_1,\, \tau_2)$	ζ	Γ	$I = 50$	$I = 100$	$I = 500$	$I = 1000$
$\left(\frac{1}{2},\, 0\right)$	9/10	1	(0.74, 0.66)	(0.86, 0.68)	(0.99, 0.68)	(1.00, 0.68)
		1.5	(0.47, 0.44)	(0.76, 0.61)	(0.99, 0.67)	(1.00, 0.67)
		2.5	(0.09, 0.08)	(0.16, 0.13)	(0.60, 0.43)	(0.87, 0.60)
		3.5	(0.02, 0.02)	(0.02, 0.01)	(0.01, 0.00)	(0.00, 0.00)
$\left(\frac{1}{2},\, 0\right)$	2/3	1	(0.81, 0.66)	(0.97, 0.68)	(1.00, 0.68)	(1.00, 0.68)
		1.5	(0.45, 0.44)	(0.77, 0.61)	(1.00, 0.67)	(1.00, 0.67)
		2.5	(0.09, 0.08)	(0.15, 0.13)	(0.49, 0.43)	(0.76, 0.60)
		3.5	(0.03, 0.02)	(0.02, 0.01)	(0.01, 0.00)	(0.00, 0.00)
$\left(\frac{1}{2},\, \frac{1}{4}\right)$	9/10	1	(0.78, 0.82)	(0.92, 0.93)	(1.00, 1.00)	(1.00, 1.00)
		1.5	(0.43, 0.48)	(0.67, 0.67)	(0.93, 0.81)	(0.98, 0.89)
		2.5	(0.08, 0.09)	(0.13, 0.13)	(0.53, 0.43)	(0.83, 0.60)
		3.5	(0.02, 0.02)	(0.01, 0.01)	(0.01, 0.00)	(0.00, 0.00)

are correct with probability $\Upsilon = \frac{2}{3}$. When $\tau_2 = 0$, the treatment has no effect on the second outcome. The design sensitivity for outcome $k = 1$ is $\widetilde{\Gamma}_1 = 3.17$ with $\tau_1 = \frac{1}{2}$; for outcome $k = 2$, $\widetilde{\Gamma}_2 = 1.76$ for $\tau_2 = \frac{1}{4}$ or $\widetilde{\Gamma}_2 = 1$ for $\tau_2 = 0$. When $\tau_1 = \frac{1}{2}$ and $\tau_2 = 0$, it is easier to identify the better outcome and more important to do so, whereas when $\tau_1 = \frac{1}{2}$ and $\tau_2 = \frac{1}{4}$, it is harder to identify the better outcome but slightly less important to do so. Two splits are considered, $\zeta = \frac{9}{10}$ and $\zeta = \frac{2}{3}$, so in both cases, most of the data is saved for use in analysis.

The limiting cases as $I \rightarrow \infty$ are consistent with the numerical values in Table 1. In all cases, by the basic property of design sensitivity, the power tends to zero as $I \rightarrow \infty$ for $\Gamma = 3.5 > 3.17 = \widetilde{\Gamma}_1 = \max\left(\widetilde{\Gamma}_1, \widetilde{\Gamma}_2\right)$. When $(\tau_1,\, \tau_2) = \left(\frac{1}{2},\, 0\right)$, the choice of outcome affects the power even in a randomization test ($\Gamma = 1$): for $I = 1000$, $\Gamma = 1$, the powers are, to two decimals, 1.00 for $\widehat{T}$ with $\zeta = \frac{9}{10}$ or $\frac{2}{3}$, and 0.68 for $\widetilde{T}$, where $0.68 = \Upsilon + (1 - \Upsilon)\, \alpha = \frac{2}{3} + \left(1 - \frac{2}{3}\right) \cdot 0.05$. In Table 1, the spitting

procedure with $\zeta = \frac{9}{10}$ does well compared to guessing correctly $\Upsilon = \frac{2}{3}$ of the time. For $I = 100$, this means using $(1-\zeta) I = 10$ observations to select the outcome and $\zeta I = 90$ observations in analysis, as opposed to guessing correctly $\Upsilon = \frac{2}{3}$ of the time and using $I = 100$ observations in analysis.

What happens if $V_{ik} \sim_{iid} N(\tau_k, 1)$, $k = 1, 2$, but with positive correlation? Positive correlation between V_{i1} and V_{i2} yields a positive correlation between the signed rank statistics in the planning sample, T_{11} and T_{21}, without altering their expectations and variances which depend only on the marginal distributions. As $I \to \infty$, $T_{11} - T_{21}$ is approximately Normal, with the same expectation as the case of independent outcomes, but with smaller variance, so the probability of selecting the correct outcome, $T_{11} - T_{21} > 0$, is increased, and the power of the split sample procedure is somewhat better than in Table 1. Negative correlation has the opposite effect.

3.3 Sample splitting versus a Bonferroni adjustment: K outcomes

Suppose that instead of two independent outcomes, as in §3.2, there are $K \geq 2$ independent outcomes, with signed rank statistics T_{k1}, T_{k2}, and T_k, for outcome k, $k = 1, \ldots, K$, in the $(1-\zeta) I = I_1$ split, the ζI split, and the full sample of size I. In the current section, outcome $k = 1$ is positively affected by treatment, with $E(T_{11}) = \mu_{11}$ and $var(T_{11}) = \sigma_{11}^2$ defined in §2.3, but outcomes $k = 2, \ldots, K$ are unaffected, so that $E(T_{k1}) = \mu_{k1} = I_1(I_1+1)/4$ and $var(T_{k1}) = \sigma_{k1}^2 = I_1(I_1+1)(2I_1+1)/24$ for $k = 2, \ldots, K$. The first outcome has the highest estimated design sensitivity if and only if $T_{11} > T_{k1}$, $k = 2, \ldots, K$, which occurs with probability approximated by $\text{PCS} = \int_{-\infty}^{\infty} \Phi^{K-1}\{(v-\mu_{k1})/\sigma_{k1}\} \ \phi\{(v-\mu_{11})/\sigma_{11}\}/\sigma_{11} \ dv$ where $\phi(\cdot)$ and $\Phi(\cdot)$

are the standard Normal density and cumulative distribution, and this formula for the probability of a correct selection (PCS) is essentially due to Bechhofer (1954).

Figure 1 plots PCS against $(1-\zeta)\,I$ for $(1-\zeta)\,I = 10, \ldots, 200$, for $K = 2$, 4, 8, 25 and 100 independent outcomes, where the treated-minus-control difference V_{ik} in pair i for outcome $k = 1$ is $N\left(\frac{1}{2}, 1\right)$ and for outcomes $k = 2, \ldots, K$ are $N(0, 1)$. In Figure 1 a planning sample of size $(1-\zeta)\,I = 50$ is sufficient to yield a high probability of selecting the correct outcome.

The top of Table 2 gives the power of the two stage procedure, in which one of the K outcomes is selected on the basis of $(1-\zeta)\,I = I_1$ observations, and the sensitivity analysis is performed for that outcome using the remaining ζI observations. The triple (0.86, 0.70, 0.54) in the upper left cell is for $K = 2$, 4, or 8 outcomes, and the power for $K = 8$ outcomes is 0.54, rather than 0.86 for $K = 2$ outcomes because it is more difficult to identify the one affected outcome when there are 8 outcomes.

As $I \to \infty$, the power is tending to a step function with a step down of size 1 at the design sensitivity for outcome $k = 1$, namely $\widetilde{\Gamma}_1 = 3.17$. This limiting behavior is seen quite clearly for $I = 500$ or $I = 1000$ in Table 2. For $I = 100$, a planning sample of size 10 for $\zeta = 9/10$ is too small, and better power is achieved with $\zeta = 2/3$, because there is a material chance of selecting the wrong outcome. For $I = 100$, the power is lower when K is higher, but for $I = 500$ or $I = 1000$, the number of outcomes, K, barely affects the power of the splitting procedure, because correct identification is highly probable, consistent with Figure 1.

The bottom of Table 2 gives the power of the Bonferroni procedure, in which rejection for outcome $k = 1$ requires a significance level less than or equal to $\alpha/K =$

Table 2: Power by Splitting When Selecting from Among 1 Affected Outcome and $K-1$ Unaffected Outcomes. The three values are power for $K = 2, 4, 8$, where the case of $K = 2$ is identical to Table 1.

$(\tau_1,\, \tau_k)$	ζ	Γ	$I = 100$	$I = 500$	$I = 1000$
$\left(\frac{1}{2},\, 0\right)$	9/10	1	(0.86, 0.70, 0.54)	(0.99, 0.98, 0.96)	(1.00, 1.00, 1.00)
		1.5	(0.76, 0.61, 0.46)	(0.99, 0.98, 0.96)	(1.00, 1.00, 1.00)
		2.5	(0.16, 0.13, 0.09)	(0.60, 0.59, 0.58)	(0.87, 0.87, 0.87)
		3.5	(0.02, 0.01, 0.01)	(0.01, 0.01, 0.01)	(0.00, 0.00, 0.00)
$\left(\frac{1}{2},\, 0\right)$	2/3	1	(0.97, 0.93, 0.88)	(1.00, 1.00, 1.00)	(1.00, 1.00, 1.00)
		1.5	(0.77, 0.74, 0.69)	(1.00, 1.00, 1.00)	(1.00, 1.00, 1.00)
		2.5	(0.15, 0.14, 0.13)	(0.49, 0.49, 0.49)	(0.76, 0.76, 0.76)
		3.5	(0.02, 0.02, 0.02)	(0.01, 0.01, 0.01)	(0.00, 0.00, 0.00)
$\left(\frac{1}{2},\, 0\right)$	Bonferroni	1	(1.00, 1.00, 0.99)	(1.00, 1.00, 1.00)	(1.00, 1.00, 1.00)
		1.5	(0.85, 0.76, 0.65)	(1.00, 1.00, 1.00)	(1.00, 1.00, 1.00)
		2.5	(0.11, 0.07, 0.04)	(0.51, 0.39, 0.28)	(0.82, 0.73, 0.63)
		3.5	(0.01, 0.00, 0.00)	(0.00, 0.00, 0.00)	(0.00, 0.00, 0.00)

$0.05/K$, using all I pairs, as described in §3.1. For a randomization test, $\Gamma = 1$ with $I = 100$ pairs, the Bonferroni procedure has higher power than split samples, which is consistent with Cox's (1975) findings, despite some differences in the models and methods. For a sensitivity analysis with $\Gamma = 2.5 < \widetilde{\Gamma}_1 = 3.17$ and $I = 500$ or $I = 1000$ pairs, the situation is reversed: with $1-\zeta = \frac{1}{10}$, split samples have higher power than the Bonferroni procedure, particularly with $K = 8$ outcomes, for instance, 0.58 versus 0.28 for $I = 500$. As one would expect from the asymptotic calculations, for both procedures, for $I = 500$ and $I = 1000$, the power is near 1 for all procedures for $\Gamma \leq 1.5 \ll \widetilde{\Gamma}_1 = 3.17$ and the power is near zero for $\Gamma = 3.5 > \widetilde{\Gamma}_1 = 3.17$.

Figure 2 plots the power for the split sample and Bonferroni procedures for several sample sizes I and several effects $(\tau_1,\, \tau_k)$. Figure 2 contains a single curve for the split sample procedure, because in this case $(1-\zeta)\, I = 100$ pairs in the planning

sample suffice to determine, with negligible probability of error, the outcome with the largest design sensitivity. As $I \rightarrow \infty$, all of the curves in Figure 2 tend to a step function with a single step down at the design sensitivity, $\widetilde{\Gamma}$, but the split sample procedure is slightly ahead, particularly for larger values of K. As emphasized in Table 2, the relative performance actually reverses for $\Gamma = 1$, but the power near $\Gamma = 1$ in Figure 2 is close to 1 in all cases. If the power is not close to 1 for $\Gamma = 1$, then the study is likely to be sensitive to small biases.

4 Split Samples and Design Sensitivity: Coherence

4.1 What is coherence?

The association between a treatment and outcomes is coherent if it is compatible with a mechanism through which the treatment is thought to produce effects. Campbell (1988, p. 33) wrote: "inferential strength is added when each theoretical parameter is exemplified in two or more ways, each mode being as independent as possible of the other, as far as the theoretically irrelevant components are concerned;" see also Hill (1965), Breslow & Day (1980, §3.2), Trochim (1985), and Reynolds & West (1987).

The *coherent signed rank test* combines signed rank tests for individual outcomes and permits a straightforward sensitivity analysis (Rosenbaum 1997). If the coherent prediction is correct, the results may be less sensitive to unobserved biases than for most or all of the component tests. In its simplest form, for several outcomes of known orientation of effect, the coherent signed rank statistic is simply the sum of the separate signed rank statistics. Instead of choosing one outcome, the coherent

signed rank statistic could use any subset of the outcomes. The planning sample determines the subset. With K outcomes, there are $2^K - 1$ subsets, or $2^8 - 1 = 255$ subsets for $K = 8$. The current section considers sample splitting in relation to coherence, with theoretical calculations in a simplified case in §4.2, and practical calculations based on simulation in §4.3.

4.2 Theoretical results for a simpler statistic

A simpler form of coherence statistic applies the usual signed rank statistic to a scalar function $y(\cdot)$ of the multivariate response $\mathbf{R}_{ij}$. The advantage of this simpler statistic is that its design sensitivity $\widetilde{\Gamma} = p_1^{'}/\left(1-p_1^{'}\right)$ follows immediately from the calculation in §2.3, and the effects of sample splitting on its design sensitivity follow immediately from considerations similar to §3.1. The disadvantage of this simpler form is that it is impractical: it is rarely if ever possible to give an a priori specification for the scalar function $y(\cdot)$, because even if one wanted to give equal weights to the K coordinates of $\mathbf{R}_{ij}$, some form of robust standardization of the coordinates would be required. The coherent signed rank test in §4.1 is practical, but the design sensitivity calculations require simulation, as developed in §4.3.

The simplest function $y(\cdot)$ is $y(\mathbf{R}_{ij}) = \boldsymbol{\lambda}^T\mathbf{R}_{ij}$ for fixed $\boldsymbol{\lambda}$, in which case $D_i = (2Z_{i1}-1)(Y_{i1}-Y_{i2}) = \boldsymbol{\lambda}^T(2Z_{i1}-1)(\mathbf{R}_{i1}-\mathbf{R}_{i2}) = \boldsymbol{\lambda}^T\mathbf{V}_i$, from which $p_1^{'}$ and $\widetilde{\Gamma} = p_1^{'}/\left(1-p_1^{'}\right)$ are determined. If $\mathbf{V}_i \sim N_K(\boldsymbol{\tau}, \boldsymbol{\Sigma})$, then $D_i \sim N\left(\boldsymbol{\lambda}^T\boldsymbol{\tau},\, \boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda}\right)$, and

$$p_1^{'} = \Pr(D_i + D_j > 0) = \Pr\left(\frac{D_i + D_j - 2\boldsymbol{\lambda}^T\boldsymbol{\tau}}{\sqrt{2\boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda}}} > \frac{-2\boldsymbol{\lambda}^T\boldsymbol{\tau}}{\sqrt{2\boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda}}}\right) = \Phi\left(\frac{\sqrt{2}\boldsymbol{\lambda}^T\boldsymbol{\tau}}{\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda}}}\right). \tag{5}$$

Table 3: Design Sensitivity with K Equally Affected, Equally Correlated Outcomes, with Effect $\tau = 0.5$ and Correlation ρ.

	$K=1$	$K=2$	$K=4$	$K=8$
$\rho = 0$	3.17	5.30	11.71	42.96
$\rho = \frac{1}{4}$	3.17	4.39	6.02	7.78
$\rho = \frac{1}{2}$	3.17	3.83	4.39	4.78
$\rho = 1$	3.17	3.17	3.17	3.17

By a familiar result (Rao 1973, 1f.1(i), p. 60), $\sup_{\boldsymbol{\lambda}} \boldsymbol{\lambda}^T\boldsymbol{\tau}/\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda}} = \boldsymbol{\tau}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}$ is attained at $\boldsymbol{\lambda} \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}$, so the design sensitivity $\widetilde{\Gamma}$ is maximized for this $\boldsymbol{\lambda}$. Insight is provided by the simple case in which $\boldsymbol{\tau} = (\tau, \ldots, \tau)^T$ and $\boldsymbol{\Sigma}$ has 1s on the diagonal and ρ off the diagonal, so $\boldsymbol{\Sigma}^{-1}\boldsymbol{\tau} \propto \boldsymbol{\lambda} = (1, \ldots, 1)^T$ and there are K equally affected outcomes, given equal weights, all with unit standard deviation and intercorrelation ρ. Then, in (5), $\sqrt{2}\boldsymbol{\lambda}^T\boldsymbol{\tau}/\sqrt{\boldsymbol{\lambda}^T\boldsymbol{\Sigma}\boldsymbol{\lambda}} = \sqrt{2}K\tau/\sqrt{K + 2\binom{K}{2}\rho}$. Table 3 gives this design sensitivity $\widetilde{\Gamma}$ for $\tau = 0.5$, $K = 1, 2, 4, 8$ outcomes and several ρ's; a similar calculation was done for the stratified rank sum statistic in Rosenbaum (2004). In Table 3, there is markedly less sensitivity to unobserved bias with $K = 8$ uncorrelated, equally affected outcomes, but the gains from coherence are reduced for $\rho = 1/2$.

Table 4 considers the bivariate case, $K = 2$, with correlation ρ and $V_{ik} \sim N(\tau_k, 1)$, $k = 1, 2$, displaying the optimal $\boldsymbol{\lambda} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\tau}$ rescaled to integer weights and the design sensitivity $\widetilde{\Gamma}$ for these weights. In Table 4, when $\rho = \frac{1}{2}$ and $(\tau_1, \tau_2) = \left(\frac{1}{2}, \frac{1}{4}\right)$, the second outcome is ignored by the best weights, $\boldsymbol{\lambda} = (1, 0)^T$, and this is also true for $\rho = 0$ and $(\tau_1, \tau_2) = \left(\frac{1}{2}, 0\right)$. Notable in Table 4 are some negative weights. For instance, $\boldsymbol{\lambda} = (4, -3)^T$ for $\rho = \frac{3}{4}$ and $(\tau_1, \tau_2) = \left(\frac{1}{2}, 0\right)$, yielding $\widetilde{\Gamma} = 6.02$ which is almost twice the design sensitivity for V_{i1} alone, namely $\widetilde{\Gamma} = 3.17$. In this case, attaching a negative weight to a correlated but unaffected outcome yields reduced

Table 4: The Optimum Weights and the Associated Design Sensitivity With Bivariate Normal Matched Pair Differences. For display, the weights are scaled to yield integer values.

(τ_1, τ_2)	$\left(\frac{1}{2}, \frac{1}{2}\right)$	$\left(\frac{1}{2}, \frac{1}{4}\right)$	$\left(\frac{1}{2}, 0\right)$
$\rho = 0$	$\boldsymbol{\lambda} = (1,1)^T$ $\widetilde{\Gamma} = 5.30$	$\boldsymbol{\lambda} = (2,1)^T$ $\widetilde{\Gamma} = 3.66$	$\boldsymbol{\lambda} = (1,0)^T$ $\widetilde{\Gamma} = 3.17$
$\rho = \frac{1}{4}$	$\boldsymbol{\lambda} = (1,1)^T$ $\widetilde{\Gamma} = 4.39$	$\boldsymbol{\lambda} = (7,2)^T$ $\widetilde{\Gamma} = 3.30$	$\boldsymbol{\lambda} = (4,-1)^T$ $\widetilde{\Gamma} = 3.30$
$\rho = \frac{1}{2}$	$\boldsymbol{\lambda} = (1,1)^T$ $\widetilde{\Gamma} = 3.83$	$\boldsymbol{\lambda} = (1,0)^T$ $\widetilde{\Gamma} = 3.17$	$\boldsymbol{\lambda} = (2,-1)^T$ $\widetilde{\Gamma} = 3.83$
$\rho = \frac{3}{4}$	$\boldsymbol{\lambda} = (1,1)^T$ $\widetilde{\Gamma} = 3.45$	$\boldsymbol{\lambda} = (5,-2)^T$ $\widetilde{\Gamma} = 3.45$	$\boldsymbol{\lambda} = (4,-3)^T$ $\widetilde{\Gamma} = 6.02$

sensitivity to unobserved bias; see Rosenbaum (1992) for related issues.

4.3 Sample Splitting Evaluated By Simulation

4.3.1 Structure of the Simulation

The simulation compares two versions of sample splitting to two feasible methods that use all of the data, and an infeasible oracle that knows with certainty some of the information we hope to discover in the first part of the split sample. For $I = 100$ and $I = 1000$ pairs, the initial sample was, respectively, 33 pairs or 100 pairs. For $I = 48$ pairs, planning samples were either 8 or 16 pairs. Results for $I = 500$ and $(1 - \zeta) = 1/10$ were similar to $I = 1000$, and are not presented. The $K = 8$ outcomes were $\mathbf{V}_i = (V_{i1}, \ldots, V_{i8})^T \sim N_8(\boldsymbol{\tau}, \boldsymbol{\Sigma})$. The five methods were as follows.

Split samples with coherence: In the planning sample, the coherent signed rank test in Rosenbaum (1997) was applied to the $2^8 - 1 = 255$ subsets of the eight

responses, and one subset was selected that produced the smallest maximum p-value for $\Gamma = 2$. That subset was used in the analysis sample.

Oracle: The oracle knew what the investigator does not, namely the true value of $\boldsymbol{\tau}$ and $\boldsymbol{\Sigma}$, and the oracle made the best choice of subset. It discarded the planning sample. One expects the oracle to be uniformly better than sample splitting, because it has, in effect, a planning sample of infinite size and an analysis sample of the same size.

Split samples selecting one outcome: This method selected the one outcome with the largest Wilcoxon signed rank statistic in the planning sample, and used that one outcome in the analysis sample. One expects it to be better than split samples with coherence when, in fact, only one outcome is affected.

Bonferroni: The Bonferroni method used all I pairs, applying the Bonferroni adjustment to the one of 8 outcomes with the smallest maximum p-value. One expects the Bonferroni method to perform well when only one outcome is affected and to perform poorly when many outcomes are affected.

Coherence: The coherence method applied the coherent signed rank test to all $K = 8$ outcomes for all I pairs. One expects the coherence method to perform well when all outcomes are strongly affected and to perform poorly when only one outcome is affected. The coherence method should be better than the oracle when all outcomes are equally affected, because the oracle performs the analysis with ζI pairs while the coherence method uses I pairs.

The simulation estimated the power of a sensitivity analysis for several values of Γ, in one thousand replications. With one thousand replications, a proportion has a

Table 5: Power of the Sensitivity Analysis for $I = 1000$ Pairs and $\Gamma = 2.5$. The planning sample is 100 of the 1000 pairs. Based on simulation of one thousand samples. Excluding the oracle, the highest power is in bold.

$\boldsymbol{\tau}$	ζI	$\left(\frac{1}{2}, 0, \ldots, 0\right)$	$\left(\frac{1}{2}, \frac{1}{2}, \frac{1}{10}, 0, \ldots, 0\right)$	$\left(\frac{1}{2}, \frac{1}{2}, \frac{2}{5}, 0, \ldots, 0\right)$	$\left(\frac{1}{4}, \frac{1}{4}, \ldots, \frac{1}{4}\right)$
Oracle	900	0.863	1.000	1.000	1.000
Split, coherence	900	0.615	**0.999**	**0.997**	0.989
Split, select one	900	**0.860**	0.861	0.750	0.000
Bonferroni	1000	0.623	0.850	0.837	0.000
Coherence	1000	0.000	0.031	0.892	**1.000**

standard error of at most $\sqrt{\frac{1}{2}\left(1 - \frac{1}{2}\right)/1000} = 0.016$. The power is the probability that the upper bound on one-sided p-value level is less than 0.05.

4.3.2 Results of the Simulation

For $I = 100$ or $I = 1000$, the upper part of Figure 3 displays the results for $\boldsymbol{\Sigma}$ equal to the identity matrix, so the eight outcomes are independent, with $\boldsymbol{\tau} = \left(\frac{1}{2}, 0, \ldots, 0\right)$ in which the best subset is $\{1\}$, $\boldsymbol{\tau} = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{10}, 0, \ldots, 0\right)$ in which the best subset is $\{1, 2\}$, $\boldsymbol{\tau} = \left(\frac{1}{2}, \frac{1}{2}, \frac{2}{5}, 0, \ldots, 0\right)$ in which the best subset is $\{1, 2, 3\}$, and $\boldsymbol{\tau} = \left(\frac{1}{4}, \frac{1}{4}, \ldots, \frac{1}{4}\right)$ in which the best subset is $\{1, 2, \ldots, 8\}$. Again, the oracle knows which subset is best. On the upper left are plots for $I = 100$ pairs, on the upper right are plots for $I = 1000$ pairs. The lower portion of Figure 3 concerns $I = 48$ pairs, with larger effects, $\boldsymbol{\tau}$, and planning samples of either 16 or 8 pairs.

Not surprisingly, the coherence method is best when the best subset is $\{1, 2, \ldots, 8\}$, for it uses this best subset and all I pairs, but it is the worst method when the best subset is $\{1\}$. Two methods select one outcome, namely the Bonferroni method and

Table 6: Power of the Sensitivity Analysis for $I = 48$ Pairs and $\Gamma = 3.5$. The planning sample is 16 of the 48 pairs. Based on simulation of one thousand samples. Excluding the oracle, the highest power is in bold.

$\boldsymbol{\tau}$	ζI	$(1, 0, \ldots, 0)$	$\left(1, 1, \frac{1}{5}, 0, \ldots, 0\right)$	$\left(1, 1, \frac{4}{5}, 0, \ldots, 0\right)$	$\left(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}\right)$
Oracle	32	0.554	0.975	0.997	0.972
Split, coherence	32	0.333	**0.704**	**0.864**	0.629
Split, select one	32	**0.540**	0.532	0.488	0.019
Bonferroni	48	0.208	0.366	0.392	0.002
Coherence	48	0.000	0.087	0.399	**1.000**

split samples using one outcome; they do well when the best subset is $\{1\}$, but in other cases they often perform poorly. For $I = 1000$, split samples with coherence is never bad; it loses in particular cases only to methods that 'know' what it tries to learn from the planning sample. Table 5 gives the estimates of power for $I = 1000$ and $\Gamma = 2.5$ where $(1 - \zeta) I = 100$ pairs are used for planning and $\zeta I = 900$ pairs are used for analysis. The Bonferroni and Coherence methods do not split and use 1000 pairs for analysis. In Table 5, split samples with coherence and the oracle are the only methods that have meaningful power in all four situations, and of course the oracle is not a feasible method.

Table 6 is similar to Table 5, except there are $I = 48$ pairs, $(1 - \zeta) I = 16$ pairs used for planning, $\Gamma = 3.5$, and the effects are larger. The pattern is qualitatively similar to Table 5, except that in Table 5 the best feasible method was close to the oracle, while in Table 6 the best feasible method is sometimes inferior to the oracle. That is, with $(1 - \zeta) I = 16$ pairs used for planning, mistakes in planning are sometimes made. For instance, in the last column of Table 6, it is best to use

Table 7: Power of the Sensitivity Analysis Using Split Samples for Coherence for $I = 1000$ Pairs with Various Patterns of Correlation. Based on simulation of one thousand samples.

Γ	1	1.5	2.5	3.5
Uncorrelated	1.000	1.000	0.974	0.689
Symmetric	1.000	0.999	0.093	0.000
Scattered Positive	1.000	1.000	0.987	0.795
Scattered Mixed	1.000	1.000	0.994	0.833

all eight outcomes in a coherent statistic. Knowing this, the oracle has power 0.972 even though it uses 32 rather than 48 observations, but split-coherence does not know this and has power 0.629.

Table 7 considers the impact of dependence among the $K = 8$ treated-minus-control differences, $\mathbf{V}_i = (V_{i1}, \ldots, V_{i8})^T$. Throughout Table 7, $\boldsymbol{\tau} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, 0, 0, 0\right)$. In all cases, the eight differences have standard deviation one, so $\boldsymbol{\Sigma}$ has ones on the diagonal. For uncorrelated outcomes, the off diagonals of $\boldsymbol{\Sigma}$ are zero, whereas for the case of symmetric correlation, the off-diagonals are $\frac{1}{2}$. In the case of scattered positive correlations, V_{ik} and $V_{i,k+4}$ have correlation $\frac{1}{2}$ for $k = 1, 2, 3, 4$ with all other correlations equal to zero. In the case of mixed scattered correlations, V_{ik} and $V_{i,k+4}$ have correlation $\frac{1}{2}$ for $k = 1, 2$, while V_{ik} and $V_{i,k+4}$ have correlation $-\frac{1}{2}$ for $k = 3, 4$ with all other correlations equal to zero. As in §4.2, the symmetrically correlated case has substantially reduced power. Although not shown in Table 7, the results for the other methods — split samples selecting one outcome, Bonferroni and coherence on all I pairs — are easy to summarize in the case of $\Gamma = 3.5$. There were four situations, three methods, and one thousand samples, so there were $4 \times 3 \times 1000 = 12,000$ opportunities to reject when $\Gamma = 3.5$, and there was only one rejection.

5 Example: Genetic Damage and Anti-Tuberculosis Drugs

To illustrate split samples in sensitivity analysis, we divide the sample in §1 into a planning sample of $(1-\zeta)\,I = \left(1-\frac{5}{6}\right)36 = 6$ observations and an analysis sample of $\zeta I = 30$ observations, and we use the planning sample to decide between a primary analysis that focuses on MN, on CA, or on their combination using the coherent signed rank statistic. In actual practice, one would use one random split. However, to examine the stability of the analysis, we repeated it for 30 independent random splits. We used the method called "split samples with coherence" in §4.3; that is, of the three choices, MN, CA or their combination, we selected the one that had the smallest maximum p-value for $\Gamma = 2$ in the planning sample. Figure 4 shows the 30 planning samples of size six. Of the 30 random splits, 19 select chromosome aberrations, 11 select the coherent statistic, and none select micronuclei.

Figure 5 shows the results for the 30 analysis samples of size $\zeta I = 30$. In each of the 30 samples, we calculate the maximum over $\mathbf{u} \in \mathcal{U}$ of all possible p-value for $\Gamma = 6$. The boxplots display the 30 maximum p-values. The four boxplots refer to micronuclei (MN), chromosome aberrations (CA), their coherent combination (Coherence), and the method selected by the planning sample (Selected). The broken lines show the three upper bounds using all $I = 36$ pairs. For the 30 analysis samples, none of the 30 upper bounds for MN is below 0.05, whereas all 90 of the other upper bounds for CA, Coherence and Selected are below 0.05. A similar pattern is seen when all $I = 36$ pairs are used. The important decision was to avoid use of MN alone, and each of the 30 samples of size 6 correctly guided that decision. Other issues, including the loss of six observations for planning, had minor effects.

6 Discussion

The design of an observational study strongly affects the degree to which its conclusions are sensitive to biases from unmeasured covariates. Aspects of design that affect sensitivity to unobserved biases include: (i) choice of a primary outcome, (ii) coherent predictions among several outcomes (Rosenbaum 1997, 2004), (iii) the pattern and magnitude of doses of treatment (Rosenbaum 2003, 2004), (iv) the strength and biases of instrumental variables (Small and Rosenbaum 2008), (v) the trade-off between the heterogeneity of experimental material and the available sample size (Rosenbaum 2005), and (vi) uncommon but dramatic treatment effects and the use of analytic strategies intended to detect them (Rosenbaum 2007). Unfortunately, it is typically difficult to make wise decisions about design in the absence of empirical data. If one performed many analyses and reported the analysis that is least sensitive to bias, then one risks capitalizing on chance and thereby substantially exaggerating the degree to which the study is insensitive to bias. In some contexts, data from earlier studies may guide design. Here, we have considered the possibility of randomly splitting the current data set into a small planning sample and a large analysis sample, where the planning sample guides decisions about design and is then discarded.

Biases addressed by a sensitivity analysis do not diminish in magnitude as the sample size, I, increases. In consequence, it may be advantageous to sacrifice a small part of the sample size in such a way that these biases are partially addressed. This is formalized using the design sensitivity, $\widetilde{\Gamma}$, which is not affected by the sacrifice of a small fraction of the sample, but which may improve with better decisions about

design. In finite samples, there is a loss of power when part of the sample is discarded, and this must be weighed against possible gains; however, in simple contexts in §3 and §4, the gains were large and the losses small.

What splitting cannot do. Splitting will not make a study insensitive to unobserved biases. Rather, if there are design decisions that would make the study less sensitive to unobserved biases, sample splitting may guide those decisions.

For splitting, the important situations are the easy situations. To say that there is some design decision that would make the study substantially less sensitive to unobserved biases is to say that the situation, when framed in the proper way, is not marginal or ambiguous. It is in such a situation, and perhaps only in such a situation, that examination of a small planning sample can provide useful guidance. What is not marginal or ambiguous has some hope of being clearly visible even in a small planning sample. In §5, the CA measure was much less sensitive to bias than the MN measure, and this was seen without much ambiguity in every one of thirty subsamples of size six. If the study would be extremely sensitive to unobserved biases no matter how it was conducted, then sample splitting has nothing to offer, but implicitly no other strategy can offer much either. Concisely: splitting won't work if nothing works, but then nothing works.

Raising $\widetilde{\Gamma}$ rather than getting closer to $\widetilde{\Gamma}$. Generally, an increase in sample size in an observational study has only a limited effect on the sensitivity of the study to unobserved biases, that limit being the design sensitivity, $\widetilde{\Gamma}$. In contrast, splitting — that is, discarding a small part of the sample size to improve the study design — holds the realistic prospect of making the study less sensitive to unob-

served biases, that is, of increasing $\widetilde{\Gamma}$. In a moderately large observational study, the sample size is being wasted if it is used only to reduce sampling variability, and not used to improve design, because reduced sampling variability has only a very limited impact on a key source of uncertainty, namely bias from unmeasured covariates. This situation in observational studies is not analogous to experiments where unobserved bias is avoided by randomization.

Exploratory uses of splitting. We have used split samples in a regimented manner. Unlike many other methods, however, splitting can be used in the planning sample in an exploratory manner to generate unanticipated insights.

Cross-validation and repeated splitting. Sections 3 and 4 evaluated the performance of a single, random split, whereas the example in §5 compared 30 distinct random splits. In the example, it was comforting to learn that, to a certain extent, the specific split did not much matter, with similar sensitivity to bias in all cases, although the split did matter in the sense that some tests were based on coherence and others on CA alone. When choices are sharply defined and made in a mechanical manner, as in §5, the use of repeated splits is an option. There is less opportunity here than in some other contexts to combine the many splits into one analysis, because the split selects the hypothesis to test, so the meaning of rejection varies from one split to the next; nonetheless, repeated splits give some indication of the stability of the result. In complex studies, the choices may be less sharply defined, and much may be gained from exploratory analysis of the planning sample; however, repeated splits are not practical in this case.

Other contexts. In the current paper, we focused on a particularly simple problem,

the best use of several outcomes in matched pairs. However, similar considerations apply in other contexts. For instance, design sensitivity for matching with multiple controls differs from pair matching only in technical details (Rosenbaum 2004). In the current paper, we have not illustrated aspects (iii) to (vi) above, but in each case there are planning decisions that (a) affect the design sensitivity, $\widetilde{\Gamma}$, (b) are difficult decisions to make without data, (c) yet the limiting design sensitivity, $\widetilde{\Gamma}$, is unchanged by sacrificing a small portion of the data to a planning sample, suggesting that sample splitting will be advantageous in moderately large samples. Would some appropriate use of available doses reduce sensitivity to unobserved biases? Would an analysis that looks for dramatic responses among a small fraction of treated subjects be less sensitive to unobserved biases than an analysis that looks for a constant effect? With some data, perhaps the data from a small planning sample, these questions can be answered with straightforward analyses.

References

Aakvik A. (2001), "Bounding a matching estimator: the case of a Norwegian training program," *Oxford Bulletin of Economics and Statistics*, 63, 115-43.

Bechhofer, R. E. (1954), "A single sample multiple decision procedure for ranking means of Normal populations with known variances," *Annals of Mathematical Statistics*, 25, 16-39.

Breslow, N. E. and Day, N. E. (1980), *The Analysis of Case-Control Studies*, Lyon: International Agency for Research on Cancer.

Campbell, D. T. (1988), *Methodology and Epistemology for Social Science*, Chicago: University of Chicago Press.

Copas, J. and Eguchi, S. (2001), "Local sensitivity approximations for selectivity bias," *Journal of the Royal Statistical Society* B 63, 871-96.

Cox, D. R. (1975), "A note on data-splitting for the evaluation of significance levels," *Biometrika*, 62, 441-4.

Cornfield, J., Haenszel, W., Hammond, E. et al. (1959), "Smoking and lung cancer," *Journal of the National Cancer Institute*," 22, 173–203.

Diprete, T. A. and Gangl, M. (2004), "Assessing bias in the estimation of causal effects," *Sociological Methodology*," 34, 271-310.

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

Hill, A. B. (1965), "The environment and disease: Association or causation?" *Proceedings of the Royal Society of Medicine*, 58, 295-300.

Imbens, G. W. (2003), "Sensitivity to exogeneity assumptions in program evaluation," *American Economic Review*, 93, 126-132.

Lehmann, E. L. (1975), *Nonparametrics*, San Francisco: Holden Day. Reprinted (1998), NJ: Prentice Hall. Reprinted (2006), New York: Springer.

Lin, D. Y., Psaty, B. M. and Kronmal, R. A. (1998), "Assessing sensitivity of regression to unmeasured confounders in observational studies," *Biometrics* 54, 948-63.

Marcus, S. M. (1997), "Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect," *Journal of Educational and Behavioral Statistics*, 22, 193-201.

Masjedi, M. R., Heidary, A., Mohammadi, F., Velayati, A. A. & Dokouhaki, P. (2000), "Chromosomal aberrations and micronuclei in lymphocytes of patients

before and after exposure to anti-tuberculosis drugs," *Mutagenesis*, 15, 489-94.

Neyman, J. (1923), "On the application of probability theory to agricultural experiments, Reprinted in *Statistical Science*, 1990, 5, 463-80.

Rao, C. R. (1973), *Linear Statistical Inference and Applications*, NY: Wiley.

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

Robins, J. M., Rotnitzky, A. and Scharfstein, D. (1999), "Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models," in *Statistical Models in Epidemiology*, Ed. E. Halloran and D. Berry, pp. 1-94, New York: Springer.

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

Rosenbaum, P. R. (1992), "Detecting bias with confidence in observational studies," *Biometrika*, 79, 367-374.

Rosenbaum, P. R. (1997), "Signed rank statistics for coherent predictions," *Biometrics*, 53, 556-566.

Rosenbaum, P. R. (1999), "Choice as an alternative to control in observational studies (with Discussion)," *Statistical Science* 14**,** 259-304.

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

Rosenbaum, P. R. (2003), "Does a dose-response relationship reduce sensitivity to hidden bias?" *Biostatistics*, 4, 1-10.

Rosenbaum, P. R. (2004), "Design sensitivity in observational studies," *Biometrika*, 91, 153-64.

Rosenbaum, P. R. (2005), "Heterogeneity and causality: Unit heterogeneity and design sensitivity in observational studies," *American Statistician*, 59, 147-152.

Rosenbaum, P. R. (2007), "Confidence intervals for uncommon but dramatic responses to treatment," *Biometrics*, 63, 1164-1171.

Rosenbaum, P. R. and Rubin, D. B. (1983), "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *Journal of the Royal Statistical Society* B, 45, 212–8.

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

Silber, J. H., Rosenbaum, P. R., Trudeau, M. E., Chen, W., Zhang, X., Lorch, S., Rapaport-Kelz, R., Mosher, R. E, Even-Shoshan, O. (2005), "Preoperative antibiotics and mortality in the elderly," *Annals of Surgery*, 242, 107-114.

Small, D. and Rosenbaum, P. R. (2008), War and wages: the strength of instrumental variables and their sensitivity to unobserved biases," *Journal of the American Statistical Association*, 103, 924-933.

Trochim, W. M. K. (1985), "Pattern matching, validity and conceptualization in program evaluation," *Evaluation Review*, 9, 575-604.

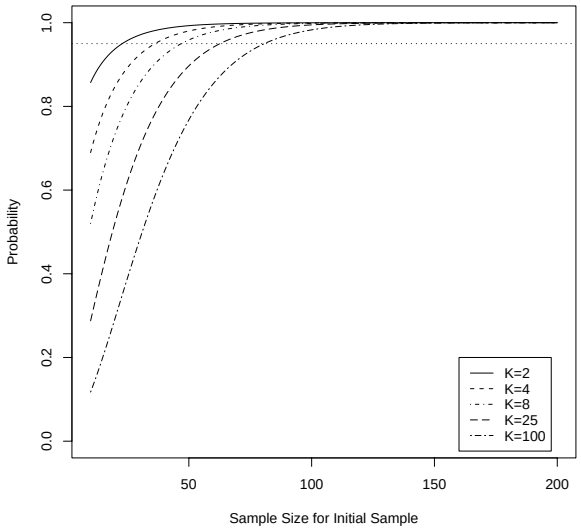

Figure 1: For various sample sizes between 10 and 200, with K independent normal outcomes, $N(\mu_k,1)$, with $\mu_1 = ½$ and $\mu_k = 0$, $k=2,\ldots,K$, the curves give the probability that the largest Wilcoxon signed rank statistic will be for outcome $k=1$. The horizontal line is at 0.95.

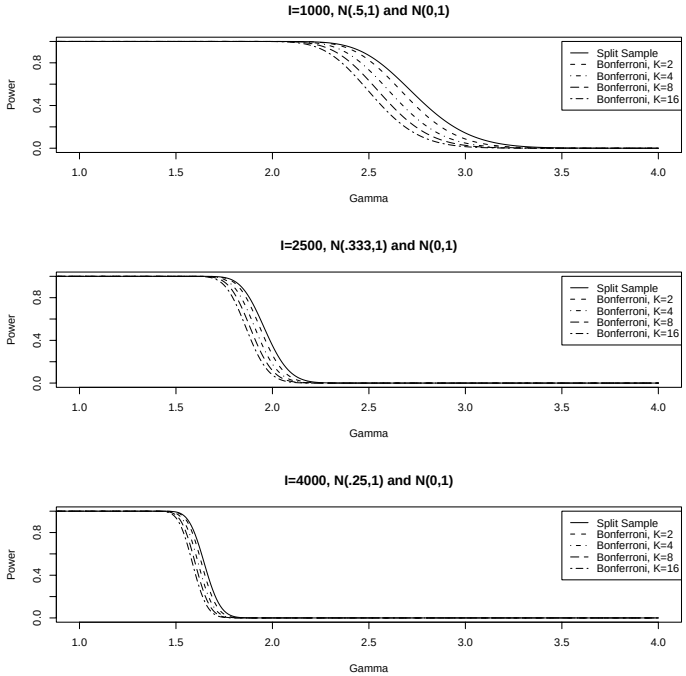

Figure 2: Power of the sensitivity analysis as a function of Γ, with $I=1000, 2500$ or 4000 pairs, $(1-\zeta)=10\%$ of pairs used in the planning sample, with $K=2, 4, 8, 16$ independent outcomes, where matched pair difference $k=1$ is $N(\tau_1, 1)$ for $\tau_1 = .5, .333, .25$, respectively, and outcomes $k=2,\ldots,K$ are $N(0,1)$. In all cases displayed, the split sample is nearly certain to identify the affected outcome, namely outcome $k=1$.

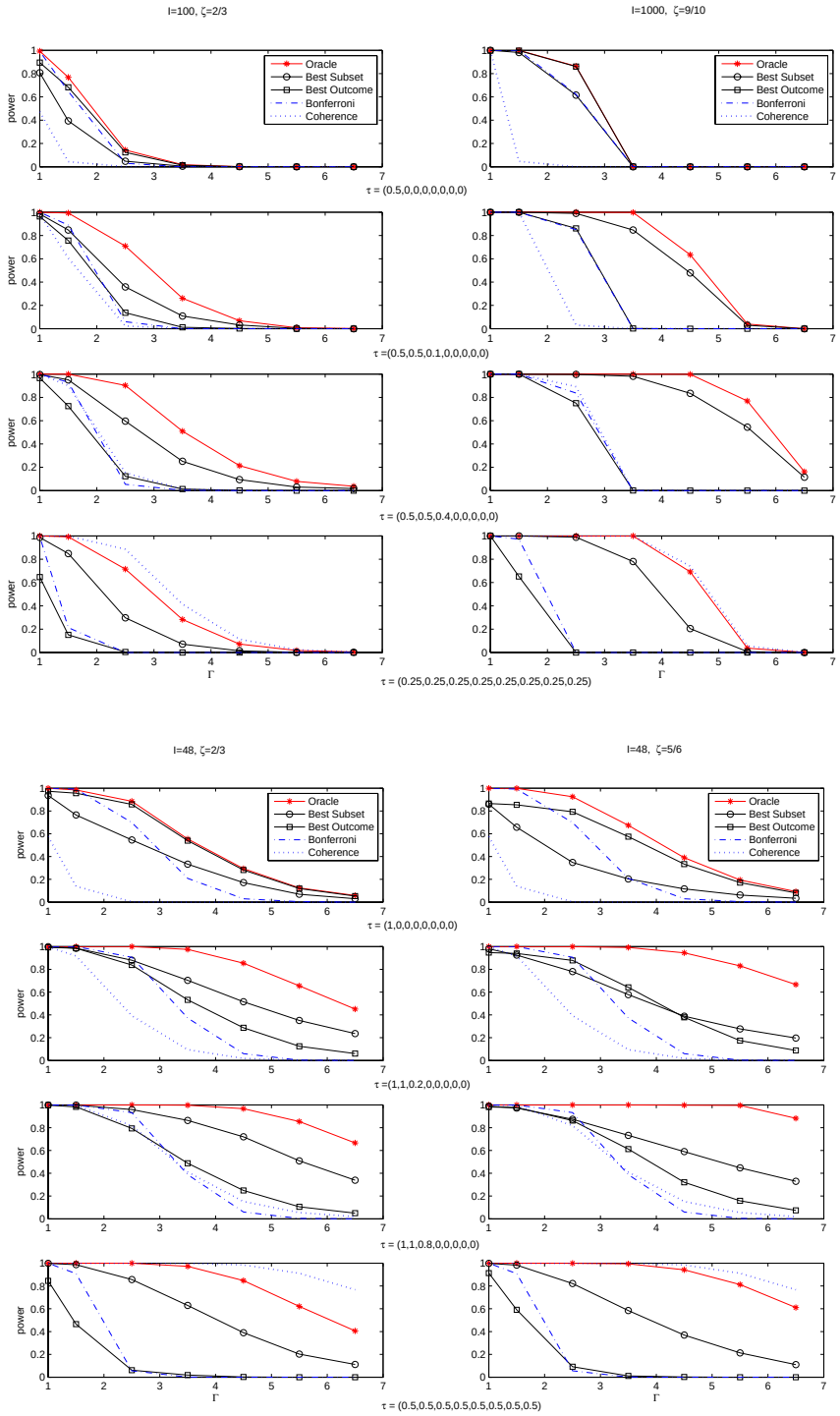

Figure 3: Power of the Sensitivity Analysis as a Function of Γ with Eight Independent Normal Differences with Expectation $\boldsymbol{\tau}$ and I Matched Pairs. Based on simulation of 1000 samples from $N_8(\boldsymbol{\tau},\mathbf{I})$. The analysis sample has ζI matched pairs.

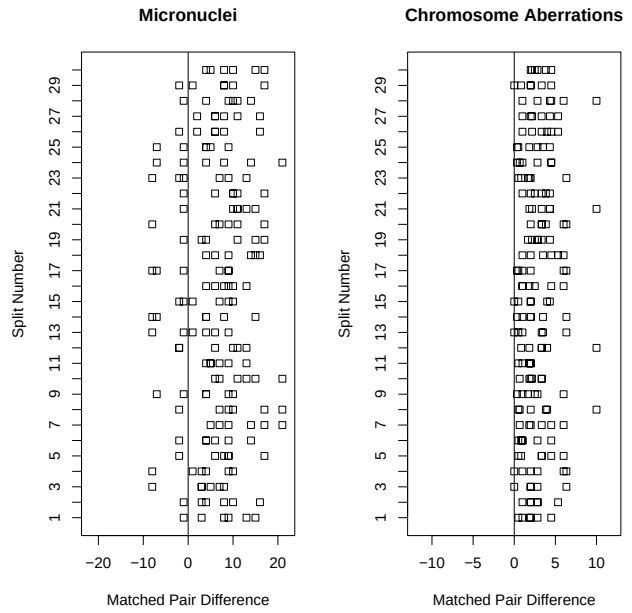

Figure 4: Micronuclei and Chromosome Aberrations in 30 Random Planning Splits of Size $(1-\zeta)I = 6$ Pairs from the $I=36$ Pairs. Of the 30 random splits, 19 select chromosome aberrations, 11 select the coherent statistic, and none select micronuclei.

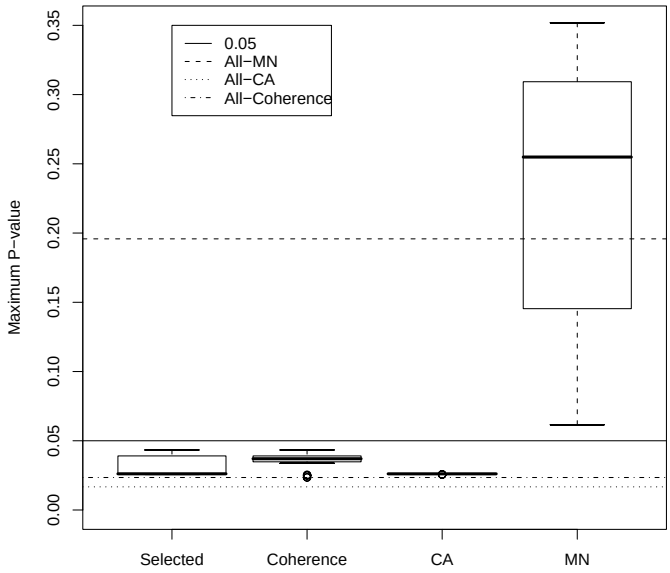

Figure 5: Upper Bounds on 30 P-values Testing No Treatment Effect with $\Gamma = 6$ for the 30 Analysis Samples of Size $\zeta I=30$. The boxplots are for micronuclei (MN), chromosome aberrations (CA), the coherent statistic (Coherence), and the method selected by the planning sample (Selected). The broken lines indicate the upper bounds on the p-values using all $I=36$ observations. The solid line is at the conventional 0.05 level. The large gain comes from not using MN alone; the other differences, including the loss of six observations to the planning sample, are small by comparison.