Biometrika (2004), **91**, 1, *pp.* 153–164

Printed in Great Britain

Design sensitivity in observational studies

BY PAUL R. ROSENBAUM

Statistics Department, Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6340, U.S.A.

rosenbaum@stat.wharton.upenn.edu

SUMMARY

Outside the field of statistics, the literature on observational studies offers advice about research designs or strategies for judging whether or not an association is causal, such as multiple operationalism or a dose-response relationship. These useful suggestions are typically informal and qualitative. A quantitative measure, design sensitivity, is proposed for measuring the contribution such strategies make in distinguishing causal effects from hidden biases. Several common strategies are then evaluated in terms of their contribution to design sensitivity. A related method for computing the power of a sensitivity analysis is also developed.

Some key words: Pattern matching; Quasi-experiment; Sensitivity analysis.

1. REDUCING SENSITIVITY BY DESIGN

1·1. *How effective are common designs for observational studies?*

Several scientific fields offer useful advice about the design of observational or non-experimental studies of treatment effects. Obviously, the first step is to adjust for observed covariates, to compare subjects who appear similar in terms of observed covariates prior to treatment, but beyond that there is invariably the concern that subjects who appear similar actually differ in terms of important unmeasured covariates. Much of the advice about design aims to reduce the 'threat to validity' from unobserved covariates. Referring to such studies as 'quasi-experiments', Cook et al. (1990, pp. 570–1) write that

> '... the warrant for causal inferences from quasi-experiments rests [on] structural elements of design other than random assignments—pretests, comparison groups, the way treatments are scheduled across groups ...—[which] provide the best way of ruling out threats to internal validity ... [C]onclusions are more plausible if they are based on evidence that corroborates numerous, complex, or numerically precise predictions drawn from a descriptive causal hypothesis.'

For representative discussions of the design of observational studies, see Yerushalmy & Palmer (1959), Hill (1965), Susser (1987), Meyer (1995) and Shadish et al. (2002). This advice is useful and widely used, but, because it is stated informally, the absolute and relative effectiveness of different strategies is not immediately apparent, and conflicts are not easily clarified.

Here, a quantitative measure, design sensitivity, is developed for appraising competing strategies in the design of observational studies. The design sensitivity is a quantity somewhat akin to Pitman efficiency: it compares the relative effectiveness of competing designs

for the same task in large samples. For a specific treatment effect and a specific research design with a large sample size, the design sensitivity asks the question of how much hidden bias would need to be present to render plausible the null hypothesis of no effect. The answer is a number, and it provides a quantitative comparison of alternative research designs. Other things being equal, we prefer the design that is less sensitive to hidden biases.

The design sensitivity is a general concept, defined in § 3, following a brief review of sensitivity analysis in § 2. The design sensitivity is then used to appraise two common design strategies, 'multiple operationalism' and the selection of treatment doses, which are reviewed in § 1·2. In § 4, the effectiveness of these two strategies is appraised and compared, and certain ostensibly conflicting claims about doses are shown, upon formalisation, to refer to different situations without conflict. This is done in a simple setting that permits exact evaluations, although the design sensitivity may be applied in other contexts as well. The effect of finite sample size and the power of a sensitivity analysis are discussed in § 5; these results are useful in planning a specific study, rather than in comparing design strategies.

1·2. *Pattern specificity: Multiple outcomes and doses*

'Succesful prediction of a complex pattern of multivariate results,' write Cook & Shadish (1994, p. 565), 'often leaves few plausible alternative explanations.' Often, only certain patterns are scientifically plausible as treatment effects (Weed & Hursting, 1998). Trochim (1985, p. 580) writes that '... with more pattern specificity it is generally less likely that plausible alternative explanations for the observed effect pattern will be forthcoming.' One form of pattern specificity is 'multiple operationalism' or 'coherence' in which several outcomes should all be affected by the treatment in a known direction. Campbell (1988, p. 33) writes that '... great 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 Reynolds & West (1987) and Li et al. (2001) for examples.

Another form of pattern specificity concerns dose-response; see Hill (1965), Weiss (1981), Susser (1987) and Rosenbaum (2003). Hill (1965, p. 298) writes that

> '... if the association is one which can reveal a biological gradient, or dose-response curve, then we should look most carefully for such evidence. For instance, the fact that the death rate from cancer of the lung rises linearly with the number of cigarettes smoked daily, adds a very great deal to the simpler evidence that cigarette smokers have a higher death rate than non-smokers.'

The available informal advice about dose-response relationships appears, at first, to be in conflict. Hill (1965) stresses that a dose-response relationship is important for causal inference. It is also said that observational studies should be patterned after simple experiments (Cochran, 1965). In experiments with human subjects, such as clinical trials, it is typically said (Peto et al., 1976, p. 590) that one should compare just two treatments that are as different as possible. Of course, these three bits of advice all seem reasonable, but also appear to conflict. If there is just a high-dose group and a zero-dose control, then the treatments are as different as possible, but there is no evidence about graduated increases in response with graduated increases in dose. It turns out however that, when these bits of advice are formalised, they are each correct in a certain sense and not in

conflict. Section 4·3 discusses the selection of doses during research design, whereas § 4·4 discusses the rather different issue of the use in analysis of whatever doses happen to be available.

2. Brief review of sensitivity analysis

2·1. *Review: Model for treatment assignment*

Although design sensitivity may be computed for a wide variety of situations, to minimise incidental technicalities, the general idea of § 3 will be illustrated in the important special case of matching with a fixed number k of controls, with $k \geqslant 1$. There are I matched sets, $i = 1, \ldots, I$, with one treated subject and k untreated controls in each matched set, $j = 1, \ldots, k+1$, where the subscript (i, j) carries no information, and the treated subject is identified by $Z_{ij} = 1$ and the controls by $Z_{ij} = 0$. The sets were matched for observed covariates, but failed to control an unobserved covariate u_{ij}. In set i, the treatment is applied at a nonnegative dose $d_i \geqslant 0$. Write $Z = (Z_{11}, Z_{12}, \ldots, Z_{I,k+1})^{\mathrm{T}}$, and write $\mathscr{Z}$ for the set containing the $(1+k)^I$ possible values of Z, so that $z \in \mathscr{Z}$ implies that z_{ij} is 1 or 0 and $1 = \sum_{j=1}^{k+1} z_{ij}$.

In a randomised experiment, one subject in each matched set would be randomly picked for treatment, the others being assigned to control, with independent assignments in the I distinct matched sets, so that $\mathrm{pr}(Z_{ij} = 1) = (k+1)^{-1}$, for each i, j. In the absence of random assignment, subjects with different values of the unobserved covariate u may have different chances of receiving the treatment. The sensitivity model assumes the following: (i) in the population before matching, treatments were assigned independently, and two subjects with the same value of the observed covariates used for matching may differ in their odds of receiving the treatment, $\mathrm{pr}(Z = 1)/\mathrm{pr}(Z = 0)$, by at most a factor of $\Gamma \geqslant 1$; (ii) subjects were exactly matched in disjoint matched sets using just observed covariates and the condition that each matched set contains one treated subject and k controls, so that $1 = \sum_{j=1}^{k+1} Z_{ij}$ for each i. It is straightforward to show (Rosenbaum, 1995; 2002, § 4.2.2) that this is exactly the same as assuming the following model, where $\gamma = \log(\Gamma) \geqslant 0$:

$$\mathrm{pr}(Z = z) = \prod_{i=1}^{I} \frac{\exp(\gamma \sum_{j=1}^{k+1} z_{ij} u_{ij})}{\sum_{j=1}^{k+1} \exp(\gamma u_{ij})}, \tag{1}$$

with $0 \leqslant u_{ij} \leqslant 1$, for all i, j, for each $z \in \mathscr{Z}$ and for some unobserved covariate u_{ij}. For discussion of unbounded u_{ij}, see Rosenbaum (1987, § 4). If $\Gamma = 1$ or $\gamma = 0$, then (1) is the randomisation distribution, $\mathrm{pr}(Z = z) = (1+k)^{-I}$. For $\Gamma > 1$, the distribution (1) is unknown because the u's are unknown, so instead of a single inference, a single significance level say, the result will be a range of significance levels, the range becoming wider as Γ increases. How large must Γ be, that is, how far must (1) depart from the randomisation distribution, to alter materially the conclusions of the study? This is the question addressed by the sensitivity analysis.

2·2. *Review: Bounds on inference for hidden biases of a given size*

Each subject exhibits a p-dimensional response, R_{ij}, with mth coordinate R_{ijm}. Under the null hypothesis of no treatment effect, the same value of R_{ij} is observed whether the subject is assigned to treatment or control, whereas alternative hypotheses assert that receiving the treatment changes a subject's observed response. Let the scalar q_{ij} be some form of rank of R_{ij}. For example, with a single response, $p = 1$, a common definition of the

rank q_{ij} entails ranking separately within each matched set from 1 to $k+1$. Pirie (1974) argues for ranking separately in different matched sets. An alternative, still with $p=1$, aligns the responses within each matched set by subtracting their mean, $R_{ij1}-\bar{R}_{i1}$, where $\bar{R}_{i1}=(k+1)^{-1}\sum_{j=1}^{k+1}R_{ij1}$, and assigns ranks q_{ij} from 1 to $I(k+1)$ to these aligned responses; see Hodges & Lehmann (1962). With $p\geqslant 2$, one common strategy calculates one of the two ranks just described from a univariate summary of R_{ij} (Dawson & Lagakos, 1993), and an alternative computes ranks for coordinates first and combines the ranks (O'Brien, 1984; Rosenbaum, 1991). Consider the statistic $T=\sum_{i=1}^{I}d_i\sum_{j=1}^{k+1}Z_{ij}q_{ij}$. For suitable k, d_i and q_{ij}, the statistic T can express a wide variety of multiple outcome summary statistics, such as a sum of p stratified Wilcoxon rank sum statistics or aligned rank statistics, a sum of p stratified Mantel & Haenszel (1959) statistics or Mantel (1963) extension statistics, or the coherent signed rank statistic (Rosenbaum, 1997). For instance, if the p responses are all binary, $R_{ijm}=1$ or $R_{ijm}=0$, and the doses are constant, $d_i=1$, then T is the sum of p correlated Mantel–Haenszel statistics with $q_{ij}=\sum_{m=1}^{p}R_{ijm}$. Sensitivity analyses for tests are inverted to yield confidence intervals and point estimates (Rosenbaum, 2002, §§ 4–5) or multivariate equivalence tests (Li et al., 2001).

A one-sided test of the hypothesis of no treatment effect rejects when T is large, and requires the computation of $\mathrm{pr}(T\geqslant k)$ under the null hypothesis. Since the u's are not observed, each Γ and each possible value of the $I(k+1)$ unobserved u's can yield a different significance level, $\mathrm{pr}(T\geqslant k)$, using (1). For several values of Γ, the sensitivity analysis computes the maximum possible value of the significance level, $\mathrm{pr}(T\geqslant k)$. For $\Gamma=1$, there is only one possible significance level, namely the usual one from the randomisation test. For each fixed $\Gamma\geqslant 1$, there is an assignment of values to the $I(k+1)$ unobserved covariates u_{ij} which provides the maximum value of $\mathrm{pr}(T\geqslant k)$, and this maximum has either $u_{ij}=0$ or $u_{ij}=1$ for each i,j with at least one 0 and one 1 in each matched set (Rosenbaum & Krieger, 1990). Gastwirth et al. (2000, § 3) give an easily computed large-sample Normal approximation to the maximum $\mathrm{pr}(T\geqslant k)$ by setting the u's to yield the maximum expectation μ_Γ of T, and, if several different patterns of u's produce the same maximum expectation, then picking from among these the one yielding the largest variance. Let $q_{i(1)}\leqslant q_{i(2)}\leqslant\ldots\leqslant q_{i(k+1)}$ be the ordered ranks for matched set i. As shown in Gastwirth et al. (2000, § 3·1), in matched set i, the largest null expectation of $\sum_{j=1}^{k+1}Z_{ij}q_{ij}$ under (1) is

$$\kappa_{\Gamma i}=\max_{a\in\{1,\ldots,k\}}\left\{\frac{\sum_{j=1}^{a}q_{i(j)}+\Gamma\sum_{j=a+1}^{k+1}q_{i(j)}}{a+\Gamma(k+1-a)}\right\}. \tag{2}$$

If the maximum in (2) is attained for each $a\in A_i\subseteq\{1,\ldots,k\}$, then find the largest variance of $\sum_{j+1}^{k+1}Z_{ij}q_{ij}$ among a's yielding this largest expectation,

$$v^2_{\Gamma i}=\max_{a\in A_i}\left\{\frac{\sum_{j=1}^{a}q^2_{i(j)}+\Gamma\sum_{j=a+1}^{k+1}q^2_{i(j)}}{a+\Gamma(k+1-a)}-\kappa^2_{\Gamma i}\right\}, \tag{3}$$

noting carefully that (3) is a maximum over A_i, not over $\{1,\ldots,k\}$. For conventional ranks, $q_{i(1)}=1,\ldots,q_{i(k+1)}=k+1$, the required values of $\kappa_{\Gamma i}$ and $v^2_{\Gamma i}$ are tabled in Gastwirth et al. (2000, Table 1). To avoid degenerate situations as $I\to\infty$, it is convenient to assume that the $v^2_{\Gamma i}$ are uniformly bounded, that is $0<v^2_{\Gamma\min}\leqslant v^2_{\Gamma i}\leqslant v^2_{\Gamma\max}$ for all i, as would automatically be true if conventional ranks $1,\ldots,k+1$ were used. Write $\mu_\Gamma=\sum d_i\kappa_{\Gamma i}$ and $\sigma^2_\Gamma=\sum d_i^2v^2_{\Gamma i}$, so that μ_Γ is the maximum expectation of T and σ^2_Γ is the maximum variance of T among patterns of the u_{ij} that yield the maximum expectation. Proposition 1 of

Gastwirth et al. (2000) shows that the maximum value of the upper tail probability under the null hypothesis, $\text{pr}(T \geqslant k)$ for $k > \mu_\Gamma$, converges to $1 - \Phi\{(k - \mu_\Gamma)/\sigma_\Gamma\}$ as $I \to \infty$, where $\Phi(.)$ is the standard Normal cumulative distribution, the approximation often being quite good for I as small as 15.

Observational studies vary considerably in their sensitivity to hidden bias. Hammond's study of heavy smoking as a cause of lung cancer is sensitive only to very large biases, $\Gamma = 6$, whereas the study of Jick et al. on coffee as a cause of myocardial infarction is sensitive to quite small biases, $\Gamma = 1{\cdot}3$; see Rosenbaum (2002, § 4). For several recent applications of this method of sensitivity analysis, see Aakvik (2001), Li et al. (2001) and Normand et al. (2001). Alternative methods of sensitivity analysis for hidden bias in observational studies are discussed by Cornfield et al. (1959), Rosenbaum & Rubin (1983), Gastwirth (1992), Copas & Li (1997), Lin et al. (1998), Robins et al. (1999), Copas & Eguchi (2001) and Imbens (2003).

As I increases, it is somewhat more convenient to work with means rather than totals, so write $T_I = T/I$, $\bar{\mu}_\Gamma = \mu_\Gamma/I$ and $\bar{\sigma}_\Gamma^2 = \sigma_\Gamma^2/I$, so that the approximate bounding Normal distribution for T_I has expectation $\bar{\mu}_\Gamma$ and variance $\bar{\sigma}_\Gamma^2/I$.

3. Design sensitivity

The definition of 'design sensitivity' applies to matching with one or more controls as in § 2, and also to other situations in which the large-sample approximation to the sensitivity bound is based on comparing a test statistic to a Normal distribution with expectation μ_Γ and variance σ_Γ^2.

Let T_I be asymptotically Normal so that

$$\text{pr}\left\{\frac{I^{\frac{1}{2}}(T_I - \mu)}{\sigma} \geqslant k\right\} \to 1 - \Phi(k) \tag{4}$$

for each fixed k, as $I \to \infty$. In (4), μ and σ^2/I are the actual expectation and variance of the limiting distribution of T_I in some situation. A large-sample statistical test of the null hypothesis of no treatment effect compares T_I not to its actual limiting distribution (4), but rather to its limiting distribution under the null hypothesis, and, in parallel, a large-sample sensitivity analysis compares T_I to its limiting bounding distribution, in typical cases a Normal distribution with expectation $\bar{\mu}_\Gamma$ and variance $\bar{\sigma}_\Gamma^2/I$. For a fixed Γ, the approximate upper bound on the one significance level is less than α if $I^{\frac{1}{2}}(T_I - \bar{\mu}_\Gamma)/\bar{\sigma}_\Gamma \geqslant k_\alpha$, where $1 - \alpha = \Phi(k_\alpha)$, and the chance that this happens satisfies

$$\begin{aligned}\text{pr}\left\{\frac{I^{\frac{1}{2}}(T_I - \bar{\mu}_\Gamma)}{\bar{\sigma}_\Gamma} \geqslant k_\alpha\right\} &= \text{pr}\left\{\frac{I^{\frac{1}{2}}(T_I - \mu)}{\sigma} \geqslant \frac{k_\alpha \bar{\sigma}_\Gamma + I^{\frac{1}{2}}(\bar{\mu}_\Gamma - \mu)}{\sigma}\right\} \\ &\to 1 - \Phi\left\{\frac{k_\alpha \bar{\sigma}_\Gamma + I^{\frac{1}{2}}(\bar{\mu}_\Gamma - \mu)}{\sigma}\right\},\end{aligned} \tag{5}$$

which tends to 1 if $\bar{\mu}_\Gamma < \mu$ and to 0 if $\bar{\mu}_\Gamma > \mu$. This says that, because hidden biases are of order $O(1)$, bias dominates the sensitivity analysis in large samples. The 'design sensitivity' is the value of Γ which solves $\bar{\mu}_\Gamma = \mu$. Since (2) and $\bar{\mu}_\Gamma$ are exact, finite-sample expectations, for the situation in § 2, the equation $\bar{\mu}_\Gamma = \mu$, and its solution, Γ, involve only exact moments, and do not directly use the Normal approximation.

In large samples, hidden bias can render the null hypothesis of no treatment effect plausible if the magnitude of hidden bias, measured by Γ, is greater than the design

sensitivity. If a first research design strategy has a larger design sensitivity than a second design, then the first design is less sensitive to bias: larger biases would have to be present to explain away the observed associations if the first design were used. Other things being equal, we would prefer a design with a larger design sensitivity. It turns out that multiple operationalism and doses affect the design sensitivity in ways that are quantified in § 4.

Without hidden bias, $\Gamma = 1$, (5) approximates the power of a randomisation test for fixed large I; see Noether (1987). For each $\Gamma \geqslant 1$, (5) approximates the probability that the maximum significance level for this Γ will be at most α; it is the analogue of power for a sensitivity analysis. Power is discussed in § 5.

4. USING DESIGN SENSITIVITY TO COMPARE DESIGNS

4·1. *A simple case: Gaussian distributions and a Wilcoxon statistic*

This section considers a simple situation that provides an informative and comprehensive yet straightforward comparison. Along the lines of Dawson & Lagakos (1983), the stratified rank sum statistic is applied to a linear summary measure, with doses and many matched sets. The model for sensitivity analysis with hidden bias and no treatment effect was given in § 2, and it yields one component of the design sensitivity, namely $\bar{\mu}_\Gamma = \mu_\Gamma / I$, using conventional ranks $q_{i(1)} = 1$, $q_{i(2)} = 2, \ldots, q_{i(k+1)} = k+1$ in (2). This is compared to a model without hidden bias, $\Gamma = 1$, so that $\text{pr}(Z = z) = (1+k)^{-I}$, but with a treatment effect. The model for the effect of the treatment on the p-variate response, R_{ij}, has a p-dimensional additive matched set parameter, α_i, a p-dimensional slope parameter, β, an effect on the treated $Z_{ij} = 1$ subject that is linear in the dose d_i with slope vector β, and independent and identically distributed p-variate continuously distributed errors, E_{ij}; that is, $R_{ij} = \alpha_i + Z_{ij}\beta d_i + E_{ij}$. A p-dimensional vector of weights, c, is selected, the scalar summary $c^{\mathrm{T}} R_{ij}$ is computed, these summaries are ranked from 1 to $k+1$ in each matched set, yielding q_{ij}, and the statistic T is computed; see Dawson & Lagakos (1993).

Let W_i be the difference between the $c^{\mathrm{T}} R_{ij}$ for the one treated subject in matched set i and the $c^{\mathrm{T}} R_{ij}$ for any one control subject in set i. In W_i, differencing has removed the matched set parameter, α_i, but has left behind the treatment effect $c^{\mathrm{T}}\beta d_i$, together with the difference of two independent errors $c^{\mathrm{T}} E_{ij}$. Now $\sum_{j=1}^{k+1} Z_{ij} q_{ij}$ is Wilcoxon's rank sum statistic with one treated subject and k controls, so it equals 1 plus the Mann–Whitney statistic, defined as the count of the number of times the treated subject had a higher $c^{\mathrm{T}} R_{ij}$ than each of the k controls, so that $\sum_{j=1}^{k+1} Z_{ij} q_{ij}$ has expectation $1 + k\,\text{pr}(W_i > 0)$; see Lehmann (1998, § 1) for the relationship between Wilcoxon's rank sum and the Mann–Whitney statistic. It follows that

$$\mu = E(T_I) = \frac{1}{I} \sum d_i \{1 + k\,\text{pr}(W_i > 0)\}. \tag{6}$$

The design sensitivity Γ solves $\bar{\mu}_\Gamma = \mu$, and this depends on the distribution of error vectors E_{ij} only through $\text{pr}(W_i > 0)$.

If the errors, E_{ij}, are p-variate Normal, $N(0, \Sigma)$, then W_i is Normal with expectation $d_i c^{\mathrm{T}}\beta$ and variance $2c^{\mathrm{T}}\Sigma c$. If we write δ_i for $(d_i c^{\mathrm{T}}\beta)/(2c^{\mathrm{T}}\Sigma c)^{\frac{1}{2}}$, it follows that $\text{pr}(W_i > 0)$ is $\Phi\{\delta_i\}$. From this, it is straightforward to solve the equation $\mu = \bar{\mu}_\Gamma$ for Γ to obtain the design sensitivity. If the errors E_{ij} are not p-variate Normal, to calculate μ one must calculate the chance that a scalar random variable, W_i, is positive. This is possible analytically for some multivariate distributions, and it is always easily done by simulation.

4·2. *Calculating the design sensitivity: A numerical illustration*

To illustrate the calculations, suppose the p-dimensional outcome is $R_{ij} = \alpha_i + Z_{ij}\beta d_i + E_{ij}$, where E_{ij} has p-variate Normal distribution with expectations equal to 0, variances equal to 1 and constant intercorrelation ρ. Suppose c is a p-dimensional vector of 1's, and the vector β has $\beta_1 = \ldots = \beta_p$, so that the responses are imperfectly correlated but have the same relationship with the dose. There are three doses, d_i, namely $\frac{1}{2}$, 1 and $\frac{3}{2}$, each occurring in one-third of matched sets. For $k = 3$ controls, $p = 3$ outcomes with correlation $\rho = \frac{1}{2}$, with each $\beta_m = \frac{1}{2}$, for $d_i = \frac{1}{2}$, 1 and $\frac{3}{2}$, one calculates, respectively, $\text{pr}(W_i > 0) = 0{\cdot}5857$, $0{\cdot}6675$ and $0{\cdot}7420$, and then $d_i\{1 + k\,\text{pr}(W_i > 0)\} = 1{\cdot}379$, $3{\cdot}002$ and $4{\cdot}839$, yielding $\mu = (1{\cdot}379 + 3{\cdot}002 + 4{\cdot}839)/3 = 3{\cdot}073$. Solving $\bar{\mu}_\Gamma = 3{\cdot}073$ gives a design sensitivity of $\Gamma = 3{\cdot}48$. For comparison, with a single outcome, $p = 1$, and other quantities as before, the design sensitivity is $\Gamma = 2{\cdot}82$, so that the use of three outcomes has reduced sensitivity to hidden bias by this magnitude.

4·3. *Appraising common strategies for design*

How effective are the suggestions in § 1·2 at reducing sensitivity to hidden bias? Table 1 offers some indications using the design sensitivity, Γ, as a measure. Table 1 allows the sample size to increase without bound, $I \to \infty$, and asks about the magnitude of hidden bias, Γ, that could explain away an observed treatment effect. The use of $p = 1$, 2 or 3 outcomes is considered, with equal correlations of $\rho = 0$ or $\rho = \frac{1}{2}$. Three dose patterns are considered. The pattern $(\frac{1}{2}, 1, \frac{3}{2})$ has three equally probable doses which average to 1; this is the only pattern which produces a dose-response relationship. The pattern (1, 1, 1) has constant dose 1. The pattern $(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$ has constant dose $\frac{3}{2}$. Comparison of $(\frac{1}{2}, 1, \frac{3}{2})$ and (1, 1, 1) indicates the value of a dose-response relationship when the average dose is the same, with or without a dose-response relationship. Comparison of $(\frac{1}{2}, 1, \frac{3}{2})$ and $(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$ contrasts the idea that a dose-response relationship is important with the competing idea that treatment and control should be as different as possible. There are $k = 2$ or $k = 5$ controls per matched set. In all cases, $\beta_1 = \ldots = \beta_p = \frac{1}{2}$, so receiving the treatment at dose 1 increases each expected response by one-half of a standard deviation.

In Table 1, the design strategies have a substantial impact on design sensitivity, which ranges from $\Gamma = 2{\cdot}15$ to $\Gamma = 11{\cdot}74$, even though the effect of the treatment at dose $d_i = 1$ is constant throughout.

Table 1. *Design sensitivity for matched studies with doses and p coherent outcomes*

		$\rho = 0$		$\rho = \frac{1}{2}$	
Doses	p	$k = 2$	$k = 5$	$k = 2$	$k = 5$
$(\frac{1}{2}, 1, \frac{3}{2})$	1	2·40	2·97	2·40	2·97
	2	3·30	4·56	2·71	3·49
	3	4·17	6·40	2·86	3·75
(1, 1, 1)	1	2·15	2·58	2·15	2·58
	2	2·86	3·75	2·39	2·97
	3	3·55	5·05	2·51	3·16
$(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$	1	3·03	4·06	3·03	4·06
	2	4·62	7·48	3·55	5·05
	3	6·37	11·74	3·81	5·59

To what extent does 'multiple operationalism' or coherence among several responses reduce sensitivity to hidden bias? There is a substantial reduction in sensitivity to hidden bias with $p = 3$ similarly affected outcomes when compared with $p = 1$ outcome providing the outcomes have uncorrelated errors, $\rho = 0$, but the gains from multiple outcomes, while still meaningful, are reduced by an intermediate correlation, $\rho = \frac{1}{2}$. For instance, with doses $(\frac{1}{2}, 1, \frac{3}{2})$ and $k = 5$ controls, the design sensitivity is $\Gamma = 6{\cdot}40$ with $p = 3$ unrelated but equally affected outcomes, but falls to $\Gamma = 3{\cdot}75$ for $p = 3$ outcomes with intercorrelation $\rho = \frac{1}{2}$, and falls further to $\Gamma = 2{\cdot}97$ with a single outcome, $p = 1$. Of course, in the limit as $\rho \to 1$, coherence is of no value.

Table 1 addresses two aspects of dose-response relationships. Consider first the comparison of doses $(\frac{1}{2}, 1, \frac{3}{2})$ with constant doses $(1, 1, 1)$, so that the average dose for treated subjects is 1 in both situations. In this comparison, a dose-response relationship does reduce sensitivity to hidden bias; for instance, with $k = 5$ controls and $p = 2$ outcomes, the design sensitivity with varied doses, $(\frac{1}{2}, 1, \frac{3}{2})$, is $\Gamma = 4{\cdot}56$, whereas with constant doses, $(1, 1, 1)$, it is $\Gamma = 3{\cdot}75$. The gain from varied doses is often of meaningful magnitude, but it is nonetheless one of the smaller sources of variation in Table 1.

The second aspect of dose-response relationships compares the varied doses $(\frac{1}{2}, 1, \frac{3}{2})$ with the fixed doses $(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$. This comparison asks whether it is better to make all of the doses as far apart as possible or to vary the doses. Here, it is quite clear that it is better to set the doses as far apart as possible rather than seek a dose-response relationship. For instance, with $k = 5$ controls and $p = 2$ outcomes, the design sensitivity with varied doses, $(\frac{1}{2}, 1, \frac{3}{2})$, is $\Gamma = 4{\cdot}56$, whereas with constant doses, $(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$, it is $\Gamma = 7{\cdot}48$.

Recall from § 1·2 that several bits of informal advice about doses appeared to conflict. Table 1 resolves this. Dose-response relationships tend to reduce sensitivity to hidden bias when compared to studies with the same average but fixed dose, but setting the doses further apart is more valuable than varying the doses to display a dose-response relationship. In observational studies, some practical approaches to setting doses further apart are discussed in Rosenbaum (1999, §§ 3.3, 3.8). A matched sampling method for picking pairs with similar covariates but very different doses is developed in Lu et al. (2001, § 2.5).

4·4. *Using or ignoring available doses in analysis*

In § 4·3 the investigator could choose the doses during research design. Suppose instead that doses simply happen to be available. Should the doses be used in analysis or ignored? Not surprisingly, the conclusion of § 4·3 is reversed: while it is better to collect data with two widely separated doses, if the data have not been collected in this way, then the analysis should use the doses that are available.

The statistic $T = \sum_{i=1}^{I} d_i \sum_{j=1}^{k+1} Z_{ij} q_{ij}$ uses the doses, attaching weight d_i to the rank sum in a matched set in which the treated subject received dose d_i, whereas the statistic $T^* = \sum_{i=1}^{I} \sum_{j=1}^{k+1} Z_{ij} q_{ij}$ ignores the doses that are present in the design, and is an unweighted sum of the I rank sum statistics. Given that doses are present in the design, how does the decision to use T or T^* affect the design sensitivity?

The design sensitivity for T, using doses, is calculated as in § 4·2, but the calculation for T^*, ignoring doses, is slightly different. For T^*, the doses do affect the responses $R_{ij} = \alpha_i + Z_{ij}\beta d_i + E_{ij}$ and so affect $\text{pr}(W_i > 0)$, but they are ignored in the statistic itself, which affects both μ and $\bar{\mu}_\Gamma$, and therefore it affects the design sensitivity. If we

continue the illustration in § 4·2 with $k=3$, $p=3$, $\rho=\frac{1}{2}$, common slope $\beta=\frac{1}{2}$ and doses $d_i=\frac{1}{2}$, 1 and $\frac{3}{2}$, the probabilities are $\text{pr}(W_i>0)=0{\cdot}5857$, 0·6675 and 0·7420, as before. Ignoring the doses using T^* yields

$$\mu=\frac{1}{I}\sum_{i=1}^{I}\{1+k\,\text{pr}(W_i>0)\}$$

which tends to $1+(0{\cdot}5857+0{\cdot}6675+0{\cdot}7420)=2{\cdot}9952$, and solving $2{\cdot}9952=\bar{\mu}_\Gamma$ yields a design sensitivity of $\Gamma=2{\cdot}96$, in contrast to $\Gamma=3{\cdot}48$ from § 4·2 for T using doses.

In Table 2, the common slope is $\beta=\frac{1}{2}$, and doses $d_i=\frac{1}{2}$, 1 and $\frac{3}{2}$ occur each with probability $\frac{1}{3}$. In all 12 situations in Table 2, using doses reduces sensitivity to hidden bias compared to ignoring doses, and the gain is substantial in a few situations, such as $k=5$, $p=3$ and $\rho=0$, where the design sensitivity improves from $\Gamma=4{\cdot}80$ to $\Gamma=6{\cdot}40$.

Table 2. *Design sensitivity using or ignoring doses present in the design with p outcomes*

		$\rho=0$		$\rho=\frac{1}{2}$	
		$k=2$	$k=5$	$k=2$	$k=5$
$p=1$	Use doses	2·40	2·97	2·40	2·97
	Ignore doses	2·13	2·56	2·13	2·56
$p=2$	Use doses	3·30	4·56	2·71	3·49
	Ignore doses	2·81	3·66	2·37	2·93
$p=3$	Use doses	4·17	6·40	2·86	3·75
	Ignore doses	3·43	4·80	2·48	3·11

Does the strength, β, of the dose-response relationship matter? Table 3 varies $\beta=\frac{1}{10}$, $\frac{1}{2}$ and 1, with $p=1$ outcome and $k=5$ controls. Here, the treatment effect at dose $d_i=1$ equals one-tenth of the standard deviation if $\beta=\frac{1}{10}$ but equals a full standard deviation if $\beta=1$. Of course, larger treatment effects, that is larger β's, yield lower design sensitivities for all designs and methods of analysis. The qualitative impression from Tables 1 and 2 is not changed by varying β: when doses happen to vary, as in rows 1 and 3, there is some improvement in design sensitivity to be had by using the doses in the analysis, particularly when the dose-response relationship is strong, $\beta=1$. Also, in rows 1 and 2, there is something to be gained by varying the doses in the design when compared to giving all treated subjects the same average dose, again particularly when the dose-response relationship is strong, $\beta=1$. Equal doses that are larger, in row 4, are best of all.

Table 3. *Doses in design versus doses in analysis: design sensitivity for four designs and analyses. (p = 1 outcome, k = 5 controls)*

Row	Doses in design	Doses in analysis	$\beta=\frac{1}{10}$	$\beta=\frac{1}{2}$	$\beta=1$
1	$(\frac{1}{2}, 1, \frac{3}{2})$	Used	1·25	2·97	8·67
2	(1, 1, 1)	Not used	1·21	2·58	6·59
3	$(\frac{1}{2}, 1, \frac{3}{2})$	Not used	1·21	2·56	6·07
4	$(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$	Not used	1·33	4·06	15·77

5. APPROXIMATE POWER FOR FIXED SAMPLE SIZES

The design sensitivity compared situations as the number of matched sets increased without bound, $I \to \infty$, and it is a concise way of comparing the relative performance of different design strategies. In planning a specific study, the sample size does matter, and, based on (5), the 'power' of the sensitivity analysis for fixed, large I is considered in this section.

Computing the approximate power of the sensitivity analysis (5) requires also the variance σ^2/I of T_I under the alternative. Consider again the situation in § 4. For conventional ranks, $q_{i(1)} = 1, \ldots, q_{i(k+1)} = k+1$, T_I is the weighted average of I Wilcoxon rank sum statistics, so the expectation is (6), as before, where W_i is the random variable

$$W_i = c^{\mathrm{T}}\beta d_i + c^{\mathrm{T}}(E_{ij} - E_{ij'}) \quad (j \neq j').$$

Write $W_i^* = c^{\mathrm{T}}\beta d_i + c^{\mathrm{T}}(E_{ij} - E_{ij''})$ with $j'' \neq j'$ and $j'' \neq j$; that is, W_i and W_i^* compare the one treated subject, j say, in matched set i to two different controls, j' and j'', in matched set i. Write $f_i = \mathrm{pr}(W_i > 0)$ and $h_i = \mathrm{pr}(W_i > 0,\ W_i^* > 0)$; then, by a standard result (Lehmann, 1998, p. 70, expression 2.21), the variance of T is

$$\sigma^2 = \sum_{i=1}^{I} d_i^2 \{k f_i(1-f_i) + k(k-1)(h_i - f_i^2)\}. \tag{7}$$

If, as in § 4, the E's are $N(0, \Sigma)$, then the (W_i, W_i^*)'s are bivariate Normal, where both coordinates have expectation $c^{\mathrm{T}}\beta d_i$ and variance $2c^{\mathrm{T}}\Sigma c$, and the covariance is $c^{\mathrm{T}}\Sigma c$. Writing $\Psi(.,.)$ for the standard bivariate Normal distribution function with correlation $\frac{1}{2}$ yields $h_i = \Psi(\delta_i, \delta_i)$ with $\delta_i = (d_i c^{\mathrm{T}}\beta)/(2c^{\mathrm{T}}\Sigma c)^{\frac{1}{2}}$.

As an illustration, consider a study with slopes $\beta_m = \frac{1}{2}$, $I = 200$ matched sets and $k = 3$ controls per set, and a one-sided, 0·05 level test with doses used in analysis when doses vary in design. If there is no hidden bias, $\Gamma = 1$, the power in all cases is nearly 1. Table 4 gives the approximate power of a sensitivity analysis for $\Gamma = 2$. The power is the chance that 0·05 is greater than the upper bound on the p-value for $\Gamma = 2$. The general pattern of the power in Table 4 for $\Gamma = 2$ is consistent with Table 1: larger constant doses $(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$ yield the highest power in Table 4, followed by varied doses $(\frac{1}{2}, 1, \frac{3}{2})$, followed by smaller constant doses (1, 1, 1), and coherence among $p = 3$ outcomes increases power, although the increase is smaller when the outcomes are correlated, $\rho = \frac{1}{2}$. If the goal were to have 80% power in a sensitivity analysis with $\Gamma = 2$ when the effect is $\beta_m = \frac{1}{2}$, then Table 4 indicates that some designs with $I = 200$ and $k = 3$ will achieve this goal while others will not.

Table 4. *Approximate power of the sensitivity analysis for $I = 200$ matched sets with $k = 3$ controls, and p outcomes*

Doses	p	$\rho = 0$	$\rho = \frac{1}{2}$
$(\frac{1}{2}, 1, \frac{3}{2})$	1	0·54	0·54
$(\frac{1}{2}, 1, \frac{3}{2})$	3	1·00	0·92
(1, 1, 1)	1	0·28	0·28
(1, 1, 1)	3	1·00	0·73
$(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$	1	0·98	0·98
$(\frac{3}{2}, \frac{3}{2}, \frac{3}{2})$	3	1·00	1·00

ACKNOWLEDGEMENT

This research was supported by a grant from the U.S. National Science Foundation.

REFERENCES

AAKVIK, A. (2001). Bounding a matching estimator: the case of a Norwegian training program. *Oxford Bull. Econ. Statist.* **63**, 115–43.

CAMPBELL, D. T. (1988). Definitional vs multiple operationalism. In *Methodology and Epistemology for Social Science*, Ed. E. S. Overman, pp. 32–6. Chicago: University of Chicago Press.

COCHRAN, W. G. (1965). The planning of observational studies of human populations (with Discussion). *J. R. Statist. Soc.* A **128**, 234–66.

COOK, T. D., CAMPBELL, D. T. & PERACCHIO, L. (1990). Quasi-experimentation. In *Handbook of Industrial and Organizational Psychology*, Ed. M. Dunnette and L. Hough, pp. 491–576. Palo Alto, CA: Consulting Psychologists Press.

COOK, T. D. & SHADISH, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Ann. Rev. Psychol.* **45**, 545–80.

COPAS, J. & EGUCHI, S. (2001). Local sensitivity approximations for selectivity bias. *J. R. Statist. Soc.* B **63**, 871–96.

COPAS, J. B. & LI, H. G. (1997). Inference for non-random samples (with Discussion). *J. R. Statist. Soc.* B **59**, 55–96.

CORNFIELD, J., HAENSZEL, W., HAMMOND, E., LILENFELD, A., SHIMKIN, M. & WYNDER, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *J. Nat. Cancer Inst.* **22**, 173–203.

DAWSON, J. D. & LAGAKOS, S. W. (1993). Size and power of two-sample tests of repeated measures data. *Biometrics* **49**, 1022–35.

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

GASTWIRTH, J. L., KRIEGER, A. M. & ROSENBAUM, P. R. (2000). Asymptotic separability in sensitivity analysis. *J. R. Statist. Soc.* B **62**, 545–55.

HILL, A. B. (1965). The environment and disease: Association or causation? *Proc. R. Soc. Med.* **58**, 295–300.

HODGES, J. L. & LEHMANN, E. L. (1962). Rank methods for combination of independent experiments in the analysis of variance. *Ann. Math. Statist.* **33**, 482–97.

IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *Am. Econ. Rev.* **93**, 126–32.

LEHMANN, E. L. (1998). *Nonparametrics.* Upper Saddle River, NJ: Prentice Hall.

LI, Y. P., PROPERT, K. J. & ROSENBAUM, P. R. (2001). Balanced risk set matching. *J. Am. Statist. Assoc.* **96**, 870–82.

LIN, D. Y., PSATY, B. M. & KRONMAL, R. A. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54**, 948–63.

LU, B., ZANUTTO, E., HORNIK, R. & ROSENBAUM, P. R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *J. Am. Statist. Assoc.* **96**, 1245–53.

MANTEL, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel–Haenszel procedure. *J. Am. Statist. Assoc.* **58**, 690–700.

MANTEL, N. & HAENSZEL, W. (1959). Statistical aspects of retrospective studies of disease. *J. Nat. Cancer Inst.* **22**, 719–48.

MEYER, B. D. (1995). Natural and quasi-experiments in economics. *J. Bus. Econ. Statist.* **13**, 151–61.

NOETHER, G. E. (1987). Sample size determination for some common nonparametric tests. *J. Am. Statist. Assoc.* **82**, 645–7.

NORMAND, S. T., LANDRUM, M. B., GUADAGNOLI, E. et al. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: A matched analysis using propensity scores. *J. Clin. Epidem.* **54**, 387–98.

O'BRIEN, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics* **40**, 1079–87.

PETO, R., PIKE, M., ARMITAGE, P., BRESLOW, N., COX, D., HOWARD, S., MANTEL, N., MCPHERSON, K., PETO, J. & SMITH, P. (1976). Design and analysis of randomised clinical trials requiring prolonged observation of each patient, I. *Br. J. Cancer* **34**, 585–612.

PIRIE, W. R. (1974). Comparing rank tests for ordered alternatives in randomised blocks. *Ann. Statist.* **2**, 374–82.

REYNOLDS, K. D. & WEST, S. G. (1987). A multiplist strategy for strengthening nonequivalent control group designs. *Eval. Rev.* **11**, 691–714.

ROBINS, J. M., ROTNITZKEY, A. & 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. (1991). Some poset statistics. *Ann. Statist.* **19**, 1091–7.

ROSENBAUM, P. R. (1995). Quantiles in nonrandom samples and observational studies. *J. Am. Statist. Assoc.* **90**, 1424–31.

ROSENBAUM, P. R. (1997). Signed rank statistics for coherent predictions. *Biometrics* **53**, 556–66.

ROSENBAUM, P. R. (1999). Choice as an alternative to control in observational studies (with Discussion). *Statist. Sci.* **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. & KRIEGER, A. M. (1990). Sensitivity analysis for two-sample permutation inferences in observational studies. *J. Am. Statist. Assoc.* **85**, 493–8.

ROSENBAUM, P. R. & RUBIN, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc.* B **45**, 212–8.

SHADISH, W. R., COOK, T. D. & CAMPBELL, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston: Houghton-Mifflin.

SUSSER, M. (1987). *Epidemiology, Health and Society: Selected Papers.* New York: Oxford University Press.

TROCHIM, W. M. K. (1985). Pattern matching, validity and conceptualization in program evaluation. *Eval. Rev.* **9**, 575–604.

WEED, D. L. & HURSTING, S. D. (1998). Biologic plausibility in causal inference: current method and practice. *Am. J. Epidem.* **147**, 415–25.

WEISS, N. (1981). Inferring causal relationships: Elaboration of the criterion of 'dose-response'. *Am. J. Epidem.* **113**, 487–90.

YERUSHALMY, J. & PALMER, C. (1959). On the methodology of investigations of etiologic factors in chronic diseases. *J. Chron. Dis.* **10**, 27–40.

[*Received July* 2002. *Revised June* 2003]