2014

Clustered Treatment Assignments and Sensitivity to Unmeasured Biases in Observational Studies

Ben B. Hansen

Paul R. Rosenbaum
University of Pennsylvania

Dylan S. Small
University of Pennsylvania

Recommended Citation

Hansen, B. B., Rosenbaum, P. R., & Small, D. S. (2014). Clustered Treatment Assignments and Sensitivity to Unmeasured Biases in Observational Studies. *Journal of the American Statistical Association, 109* (505), 133-144. http://dx.doi.org/10.1080/01621459.2013.863157

Clustered Treatment Assignments and Sensitivity to Unmeasured Biases in Observational Studies

Abstract

Clustered treatment assignment occurs when individuals are grouped into clusters prior to treatment and whole clusters, not individuals, are assigned to treatment or control. In randomized trials, clustered assignments may be required because the treatment must be applied to all children in a classroom, or to all patients at a clinic, or to all radio listeners in the same media market. The most common cluster randomized design pairs 2*S* clusters into *S* pairs based on similar pretreatment covariates, then picks one cluster in each pair at random for treatment, the other cluster being assigned to control. Typically, group randomization increases sampling variability and so is less efficient, less powerful, than randomization at the individual level, but it may be unavoidable when it is impractical to treat just a few people within each cluster. Related issues arise in nonrandomized, observational studies of treatment effects, but in this case one must examine the sensitivity of conclusions to bias from nonrandom selection of clusters for treatment. Although clustered assignment increases sampling variability in observational studies, as it does in randomized experiments, it also tends to decrease sensitivity to unmeasured biases, and as the number of cluster pairs increases the latter effect overtakes the former, dominating it when allowance is made for nontrivial biases in treatment assignment. Intuitively, a given magnitude of departure from random assignment can do more harm if it acts on individual students than if it is restricted to act on whole classes, because the bias is unable to pick the strongest individual students for treatment, and this is especially true if a serious effort is made to pair clusters that appeared similar prior to treatment. We examine this issue using an asymptotic measure, the design sensitivity, some inequalities that exploit convexity, simulation, and an application concerned with the flooding of villages in Bangladesh.

Keywords

community intervention, design sensitivity, group randomization, multi-level study, quasi-experiment, sensitivity analysis

Disciplines

Statistics and Probability

Clustered Treatment Assignments and Sensitivity to Unmeasured Biases in Observational Studies

Ben B. Hansen, Paul R. Rosenbaum, and Dylan S. Small[1]

Abstract. Clustered treatment assignment occurs when individuals are grouped into clusters prior to treatment and whole clusters, not individuals, are assigned to treatment or control. In randomized trials, clustered assignments may be required because the treatment must be applied to all children in a classroom, or to all patients at a clinic, or to all radio listeners in the same media market. The most common cluster randomized design pairs $2S$ clusters into S pairs based on similar pretreatment covariates, then picks one cluster in each pair at random for treatment, the other cluster being assigned to control. Typically, group randomization increases sampling variability and so is less efficient, less powerful, than randomization at the individual level, but it may be unavoidable when it is impractical to treat just a few people within each cluster. Related issues arise in nonrandomized, observational studies of treatment effects, but in this case one must examine the sensitivity of conclusions to bias from nonrandom selection of clusters for treatment. Although clustered assignment increases sampling variability in observational studies, as it does in randomized experiments, it also tends to decrease sensitivity to unmeasured biases, and as the number of cluster pairs increases the latter effect overtakes the former, dominating it when allowance is made for nontrivial biases in treatment assignment. Intuitively, a given magnitude of departure from random assignment can do more harm if it acts on individual students than if it is restricted to act on whole classes, because the bias is unable to pick the strongest individual students for treatment, and this is especially true if a serious

[1] Ben B. Hansen is Associate Professor, Department of Statistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: bbh@umich.edu). Paul R. Rosenbaum (E-mail: rosenbaum@wharton.upenn.edu) is Professor and Dylan S. Small (E-mail: dsmall@wharton.upenn.edu) is Associate Professor, Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104. This study was supported by grants SES-0753164 and SES-1260782 from the Measurement, Methodology and Statistics Program of the U.S. National Science Foundation. The authors acknowledge very helpful comments from three referees, an associate editor, David Silver and Joseph Ibrahim.

effort is made to pair clusters that appeared similar prior to treatment. We examine this issue using an asymptotic measure, the design sensitivity, some inequalities that exploit convexity, simulation, and an application concerned with the flooding of villages in Bangladesh.

Keywords: Design sensitivity; group randomization; sensitivity analysis.

1 Introduction; motivating example; outline

1.1 Clustered experiments and observational studies

Some treatments can be applied to a cluster of individuals but not to a single individual. For instance, the Prospect Randomized Trial (Bruce, Ten Have, Reynolds et al. 2004, Small, Ten Have and Rosenbaum 2008) paired 20 medical practices into 10 pairs of two practices so that paired practices were similar, then selected one practice in each pair at random to receive a "depression care manager" – a psychiatric nurse with special training – who provided depression-related services to patients at that practice and depression-related guidance to physicians at that practice. Similarly, Hansen and Bowers (2009) discuss the effects of a randomized get-out-the-vote campaign that could not be applied at the individual level. The same situation arises when a treatment must be applied or withheld from a school rather than from individual student, or when a public health campaign must be applied to a community rather than to individuals within that community.

In randomized experiments, clustered treatment assignment may be necessary, but it tends to reduce efficiency compared to assignment at the individual level, particularly when individuals in the same cluster tend to exhibit similar responses for reasons unrelated to the treatment (Cornfield 1978, Murray 1998).

In nonrandomized studies of treatment effects, efficiency is a secondary concern, and biases from nonrandomized treatment assignment are the primary concern (Cochran 1965). To some extent, biases from nonrandom assignment can be removed by adjustments for measured covariates, for instance, by matching or covariance adjustment. However, the concern is invariably raised that individuals or clusters that appear similar in measured covariates may differ in ways not measured, so adjustments for measured covariates may fail to compare comparable units under alternative treatments. A sensitivity analysis asks about the magnitude of the departure from random assignment that would need to be present to alter the conclusions of a naive analysis that assumes adjustments for measured covariates suffice to remove all bias. The power of a sensitivity analysis and the design sensitivity anticipate the outcome of a sensitivity analysis under an assumed model for the generation of the data, and in this sense they parallel and generalize the power of a test in a randomized experiment.

As demonstrated in the current paper, clustered treatment assignments are less susceptible to biases from unmeasured covariates than are assignments at the individual level. At an intuitive level, a bias of a given magnitude in treatment assignment can do more harm if it can pick and choose among individuals, and does somewhat less harm when forced to make the more constrained choice of picking and choosing among clusters of individuals. If a depression-care manager focused her attention on the most depressed patients then the biases could be much larger than if she elected to work at a medical practice whose patients tended to be more depressed.

1.2 Motivating example: Flooding in Bangladesh

In 1998, parts of Bangladesh experienced massive floods, while other areas were spared. Del Ninno, Dorosh, Smith and Roy (2001) conducted an observational study of the effects of flooding on health and other outcomes. We use their data to illustrate issues that arise in observational studies with clustered treatment assignments. Massive floods affect or spare villages, not individuals.

Table 1 describes 27 pairs of two villages in Bangladesh, one severely flooded, the other not exposed to the flood. Within each village, a small number of children were sampled and covariates and outcomes describe these children. In total, there were 291 children.

The outcome is the number of sick days in the two weeks following the flood. The villages were paired using three covariates: the proportion of boys among the sampled children, the mean age of the sampled children, and the median preflood assets of their families. The pairing was based on a rank-based Mahalanobis distance and the optimal assignment algorithm as implemented in the `pairmatch` function of the `optmatch` package in `R`; see Hansen (2007) or Stuart (2010). Additional adjustments will be made later by covariance adjustment for differences among the 291 children. The general impression in Table 1 is that children in flooded villages had more sick days than children in villages not exposed to the flood.

A group randomized experiment would have treated one village picked at random within each pair, but obviously, villages were not selected for flooding at random. Because villages were flooded, the deviations from random assignment affect whole villages: the nonrandom assignment cannot pick and choose for flooding among children in the same village.

2 Treatments assigned to paired clusters

2.1 Clusters matched for covariates

There are S strata or pairs, $s = 1, \ldots, S$, of two clusters, $k = 1, 2$, so the ordered pair (s, k) (or briefly sk) identifies a unique cluster. In Table 1, there are $S = 27$ pairs of two villages. Cluster sk contains $n_{sk} \geq 1$ individuals, $i = 1, \ldots, n_{sk}$. A covariate is a variable whose value is determined prior to treatment assignment and hence is unaltered when treatments are assigned. Individual i in cluster sk is described by an observed covariate $\mathbf{x}_{ski}$ and an unobserved covariate u_{ski}. The covariate $(\mathbf{x}_{ski}, u_{ski})$ may describe the individual ski and/or the cluster sk containing this individual and/or the stratum s containing this pair of clusters. In the example, there are six covariates, the child's age and gender, the child's family's preflood assets, the proportion of boys in the village sample, the mean age in the village sample and the median of preflood assets in the village sample. Whole clusters are assigned to treatment, denoted $Z_{sk} = 1$, or to control, denoted $Z_{sk} = 0$, where each pair contains one treated and one control cluster, $1 = Z_{s1} + Z_{s2}$ for each s. The pairs of clusters are typically formed by matching for observed covariates $\mathbf{x}_{ski}$ describing the clusters and the individuals within the clusters, as was done in Table 1.

Write $\mathbf{Z} = (Z_{11}, \ldots, Z_{S2})^T$ for the treatment assignments for all $2S$ clusters. If $\mathcal{S}$ is a finite set, write $|\mathcal{S}|$ for the number of elements of $\mathcal{S}$. Write $\mathcal{Z}$ for the set of possible values $\mathbf{z}$ of $\mathbf{Z}$, so $\mathbf{z} \in \mathcal{Z}$ if $\mathbf{z} = (z_{11}, \ldots, z_{S2})^T$ with $z_{s1} + z_{s2} = 1$ for $s = 1, \ldots, S$, and $|\mathcal{Z}| = 2^S$. Conditioning on the event $\mathbf{Z} \in \mathcal{Z}$ is abbreviated as conditioning on $\mathcal{Z}$. If $n_{sk} = 1$ for all sk, then the clusters are individuals, so there is no need for a separate notation for studies with unclustered treatment assignment.

2.2 Responses of individuals when whole clusters are assigned to treatment

Each individual ski has two potential responses, namely response r_{Tski} if cluster sk is assigned to treatment, $Z_{sk} = 1$, or response r_{Cski} if cluster sk is assigned to control, $Z_{sk} = 0$. There is no presumption here that individuals within the same cluster do not interfere with one another; rather, r_{Tski} describes the response of ski if *all* individuals in cluster sk receive the treatment, $Z_{sk} = 1$, and r_{Cski} describes the response of individual ski if *all* individuals in cluster sk receive the control, $Z_{sk} = 0$, and there is no presumption that these same responses would be seen from ski if treatments were assigned to some but not all individuals in cluster sk. Because each cluster receives either treatment or control, either r_{Tski} is observed or r_{Cski} is observed but never both — that is, $R_{ski} = Z_{sk}\, r_{Tski} + (1 - Z_{sk})\, r_{Cski}$ is observed — and the effect on individual ski of treating cluster sk, namely $r_{Tski} - r_{Cski}$, is not observed for any individual, in parallel with the situation without clusters described by Neyman (1923), Welch (1937) and Rubin (1974). Here, the observed response R_{ski}

changes when treatment changes Z_{sk} if $r_{Tski} - r_{Cski} \neq 0$, but (r_{Tski}, r_{Cski}) does not change as Z_{sk} changes.

In the flood example, r_{Tski} is the number of sick days that child ski would exhibit if her village were severely flooded and r_{Cski} is the number of days this same child would exhibit if her village were not exposed to the flood. The effect $r_{Tski} - r_{Cski}$ on the sick days of child ski of severe flooding of her village could in part reflect a shortage of clean water and overwhelming of medical staff in her village. Quite plausibly, the flooding of just her house but not the village would have had a very different effect on her, because then clean water and medical staff would not have been in short supply. Because the flood affected regions and not isolated homes, the available data speak to the issue of the effects of flooding of villages, not the effects of flooding of individual homes in otherwise dry villages. See Small et al. (2008) for discussion of treatment effects $r_{Tski} - r_{Cski}$ at the individual level when whole clusters are assigned to treatment or control.

In the Bangladesh example, part of the treatment effect may be produced by overwhelming the village's community services, so the effect of flooding on an individual may reflect the presence of many individuals experiencing flooding at the same time. There are other contexts in which it is convenient to assign treatment or control to whole clusters, but the effect of the treatment on an individual does not depend upon the treatments received by other individuals. Cox (1952, §2.4) refers to this as "no interference between units." Typically, an antihypertensive drug affects only the person who receives it, and in this case there is no interference between units, whether treatments are assigned to individuals or clusters. When there is no interference between units, the investigator has a choice of study designs, clustered or individual treatment assignment, but the effect caused by the treatment is the same. When an investigator can study the same effect in two different ways, it is of interest to know whether one design has advantages over the other.

Fisher's (1935) sharp null hypothesis of no treatment effect asserts that changing the treatment assigned to cluster sk would leave the response of individual ski unchanged for all individuals ski, that is, $H_0 : r_{Tski} = r_{Cski}$, $\forall ski$. Write $\mathbf{r}_C = (r_{C111}, \ldots, r_{CS2,n_{S2}})^T$ for the $N = \sum_{s,k} n_{sk}$ dimensional vector, with a similar notation for $\mathbf{r}_T$, $\mathbf{R}$, $\mathbf{u}$, etc. If Fisher's H_0 were true, $R_{ski} = r_{Cski}$ for all ski or $\mathbf{R} = \mathbf{r}_C$. Write

$$\mathcal{F} = \left\{ (r_{Tski}, r_{Cski}, \mathbf{x}_{ski}, u_{ski}) \, , \; i = 1, \ldots, n_{sk}, \; s = 1, \ldots, S, \; k = 1, 2 \right\},$$

noting that, unlike $\mathbf{R}$, the quantities in $\mathcal{F}$ are fixed, not changing as $\mathbf{Z}$ changes.

2.3 Random assignment of treatment to clusters; randomization inference

To say that treatment assignment is randomly assigned to clusters is to say that random numbers were used in the assignment of treatment in such a way that $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\mathcal{Z}| = 1/2^S$ for each $\mathbf{z} \in \mathcal{Z}$; equivalently, $Z_{s2} = 1 - Z_{s1}$, the Z_{s1} are independent for distinct s, and $\Pr(Z_{s1} = 1 \mid \mathcal{F}, \mathcal{Z}) = 1/2$ for every s.

A test statistic T is a function of $\mathbf{Z}$ and $\mathbf{R}$, that is, $T = t(\mathbf{Z}, \mathbf{R})$. If the null hypothesis H_0 were true then $\mathbf{R} = \mathbf{r}_C$, so $T = t(\mathbf{Z}, \mathbf{r}_C)$. If $\mathbf{Z}$ were randomly assigned, then the randomization distribution of T under the null hypothesis H_0 would be:

$$\Pr\{t(\mathbf{Z}, \mathbf{R}) \geq c \mid \mathcal{F}, \mathcal{Z}\} = \Pr\{t(\mathbf{Z}, \mathbf{r}_C) \geq c \mid \mathcal{F}, \mathcal{Z}\} = \frac{|\{\mathbf{z} \in \mathcal{Z} : t(\mathbf{Z}, \mathbf{r}_C) \geq c\}|}{|\mathcal{Z}|}, \tag{1}$$

because $\mathbf{r}_C$ is fixed by conditioning upon $\mathcal{F}$, and $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\mathcal{Z}|$.

Let q_{ski} be a score or rank given to R_{ski}, so that under H_0 the q_{ski} are functions of the r_{Cski} and $\mathbf{x}_{ski}$, and they do not vary with Z_{sk}. Taking $q_{ski} = R_{ski}$ yields the randomization distribution of the mean or the so-called "permutational t-test," as discussed by Pitman (1937) and Welch (1937). In practice, it will often be appropriate to stabilize R_{ski} through covariance adjustment for $\mathbf{x}_{ski}$ and to use scores q_{ski} resistant to outliers. In the example, as in Small et al. (2008), the q_{ski} in Table 1 are ranks of the residuals of R_{ski} when regressed on the six covariates in $\mathbf{x}_{ski}$ using Huber's m-estimation (with the default settings in R); see Rosenbaum (2002a) for discussion of covariance adjustment of permutation tests as well as pivoting to produce point estimates and confidence intervals.

Under H_0, the observed response R_{ski} equals r_{Cski}, and r_{Cski} is in $\mathcal{F}$, so under H_0 the ranks q_{ski} are fixed by conditioning on $\mathcal{F}$ in (1); hence, also, the mean rank $n_{sk}^{-1} \sum_i q_{ski}$ in cluster sk is fixed, not changing with Z_{sk}. Consider as a test statistic T a weighted sum over the S pairs of the mean rank in the treated cluster ($Z_{sk} = 1$) minus the mean rank in the control cluster ($Z_{sk} = 0$ or $1 - Z_{sk} = 1$), where the weight $w_s \geq 0$ for pair s is a function of the n_{sk}. Under H_0, using $Z_{s2} = 1 - Z_{s1}$, the statistic T is

$$\begin{aligned} T &= \sum_{s=1}^{S} w_s Z_{s1} \left(\frac{1}{n_{s1}} \sum_{i=1}^{n_{s1}} q_{s1i} - \frac{1}{n_{s2}} \sum_{i=1}^{n_{s2}} q_{s2i} \right) + w_s Z_{s2} \left(\frac{1}{n_{s2}} \sum_{i=1}^{n_{s2}} q_{s2i} - \frac{1}{n_{s1}} \sum_{i=1}^{n_{s1}} q_{s1i} \right) \quad (2) \\ &= \sum_{s=1}^{S} w_s \left(2Z_{s1} - 1 \right) \left(\frac{1}{n_{s1}} \sum_{i=1}^{n_{s1}} q_{s1i} - \frac{1}{n_{s2}} \sum_{i=1}^{n_{s2}} q_{s2i} \right) = \sum_{s=1}^{S} B_s \, Q_s \end{aligned}$$

where

$$B_s = 2Z_{s1} - 1 = \pm 1, \qquad Q_s = \frac{w_s}{n_{s1}} \sum_{i=1}^{n_{s1}} q_{s1i} - \frac{w_s}{n_{s2}} \sum_{i=1}^{n_{s2}} q_{s2i}. \qquad (3)$$

In (1) in a cluster randomized experiment, under H_0 given $\mathcal{F}$, $\mathcal{Z}$, the statistic T in (2) is the sum of S independent random variables taking the value $\pm Q_s$ each with probability 1/2, so $\mathrm{E}(T) = 0$ and $\mathrm{var}(T) = \sum_{s=1}^{S} Q_s^2$. Under H_0 in a group randomized experiment, for reasonable ranks, q_{ski}, as $S \to \infty$ with n_{sk} bounded, $1 \leq n_{sk} \leq \nu$, the central limit theorem implies $T/\sqrt{\mathrm{var}(T)}$ converges in distribution to the standard Normal distribution, $\Phi(\cdot)$.

Because of its analytical simplicity, several results that we present will concern the “permutational t-test” which uses the responses directly, $q_{ski} = R_{ski}$, so that Q_s is proportional to difference in mean responses in two paired clusters, and T is the weighted sum over pairs s of the treated-minus-control difference in mean responses. See Pitman (1937) and Welch (1937) for discussion of randomization inference with $q_{ski} = R_{ski}$. Other results will concern ranks calculated separately within each pair of clusters, so that T is linearly related to a weighted combination of Wilcoxon rank sum statistics (e.g., van Elteren 1960, Lehmann 1975, §3.3). In simulations, statistics that rank across clusters are also considered; see, for instance, Mantel (1977), Conover and Iman (1981) and Lam and Longnecker (1983).

In a randomized experiment, the analysis described in the current section is the same as the analysis proposed by Small, Ten Have and Rosenbaum (2008). If this randomization test is applied to the data in Table 1 with equal weights $w_s = 1$, then an approximate one-sided P-value of 0.0064 is obtained, rejecting H_0 in favor of greater illness in flooded villages. Of course, Table 1 is not from a randomized experiment.

2.4 Biased assignment of treatments to clusters; sensitivity analysis

In a nonrandomized observational study, there is nothing to ensure $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\mathcal{Z}|$, and treatment assignments may exhibit systematic biases; for instance, $\Pr(Z_{sk} = 1 \mid \mathcal{F}, \mathcal{Z})$ might vary with the unobserved covariates u_{ski} describing individuals in a cluster. To say that the assignment of treatments to clusters may be biased after matching clusters for observed covariates is to say that $\Pr(Z_{sk} = 1 \mid \mathcal{F}, \mathcal{Z})$ may deviate from 1/2 because $\Pr(Z_{sk} = 1 \mid \mathcal{F}, \mathcal{Z})$ is varying with elements of $\mathcal{F}$ that were not controlled by the matching, that is, typically, the elements of $\mathcal{F}$ that were not observed.

The possible impact of biases of various magnitudes in assignment of treatments to clusters is examined using a sensitivity analysis model that asserts the Z_{s1} are independent for distinct s with

$$\frac{1}{1+\Gamma} \leq \Pr\left(Z_{s1}=1 \mid \mathcal{F},\, \mathcal{Z}\right) \leq \frac{\Gamma}{1+\Gamma}, \quad Z_{s2}=1-Z_{s1}, \tag{4}$$

for each s, where $\Gamma \geq 1$ is a sensitivity parameter whose value is varied to examine the degree of sensitivity of conclusions to unmeasured biases. In words, (4) allows $\Pr\left(Z_{s1}=1 \mid \mathcal{F},\, \mathcal{Z}\right)$ and $\Pr\left(Z_{s2}=1 \mid \mathcal{F},\, \mathcal{Z}\right)$ to differ by at most a factor of Γ, so (4) introduces a bias in treatment assignment whose magnitude is controlled by the value of Γ. For treatment assignment at the individual level, the model (4) was proposed in Rosenbaum (1987), and various generalizations and alternative descriptions of the this model are developed in Rosenbaum (2002b, §4). Using Wolfe's (1974) semiparametric family of deformations of a symmetric distribution, Rosenbaum and Silber (2009) interpret Γ in terms of two parameters, one connecting u_{ski} with treatment assignment, the other connecting u_{ski} with outcomes. For alternative models for sensitivity analysis in observational studies, see Cornfield et al. (1959), Copas and Eguchi (2001), Gastwirth (1992), Hosman, Hansen and Holland (2010), Imbens (2003), Marcus (1997), Rosenbaum and Rubin (1983), Small (2007) and Yu and Gastwirth (2005).

Let $\theta=\Gamma/\left(1+\Gamma\right)$ and define $\overline{\pi}_s=\theta$ if $Q_s>0$ and $\overline{\pi}_s=1-\theta$ otherwise, and define $\tilde{\pi}_s=1-\overline{\pi}_s$. Let $\overline{T}_\Gamma$ be a random variable formed as the sum of S independent random variables taking the value Q_s with probability $\overline{\pi}_s$ and the value $-Q_s$ with probability $1-\overline{\pi}_s$, and let $\tilde{T}_\Gamma$ be defined in the same way but with $\tilde{\pi}_s$ in place of $\overline{\pi}_s$. Then it is not difficult to show (Rosenbaum 1987; 2002b, §4) that (4) implies

$$\Pr\left(\tilde{T}_\Gamma \geq t \,\middle|\, \mathcal{F},\, \mathcal{Z}\right) \leq \Pr\left(T \geq t \mid \mathcal{F},\, \mathcal{Z}\right) \leq \Pr\left(\overline{T}_\Gamma \geq t \mid \mathcal{F},\, \mathcal{Z}\right) \text{ for each } t. \tag{5}$$

For large S, the distribution of $\overline{T}_\Gamma$ in (5) may be approximated by a Normal distribution with expectation

$$\mathrm{E}\left(\overline{T}_{\Gamma} \mid \mathcal{F},\ \mathcal{Z}\right) = \sum_{s=1}^{S} \left(2\overline{\pi}_s - 1\right) Q_s = \frac{\Gamma - 1}{\Gamma + 1} \sum_{s=1}^{S} |Q_s|$$

and variance

$$\mathrm{var}\left(\overline{T}_{\Gamma} \mid \mathcal{F},\ \mathcal{Z}\right) = 4 \sum_{s=1}^{S} \overline{\pi}_s \left(1 - \overline{\pi}_s\right) Q_s^2 = \frac{4\,\Gamma}{\left(1+\Gamma\right)^2} \sum_{s=1}^{S} Q_s^2$$

so the upper bound on the approximate one-sided P -value is less than or equal to α if

$$\frac{T/S - \left[\left(\Gamma - 1\right) / \left\{S\left(\Gamma+1\right)\right\}\right] \sum_{s=1}^{S} |Q_s|}{\sqrt{\left[4\Gamma / \left\{S^2 \left(1+\Gamma\right)^2\right\}\right] \sum_{s=1}^{S} Q_s^2}} \geq \Phi^{-1}\left(1 - \alpha\right), \qquad (6)$$

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

If each cluster contains a single individual, $n_{sk} = 1$ for all sk, then the analysis described in §2.4 is the same as the analysis in Rosenbaum (1987; 2002b, §4). For $n_{sk} \geq 1$ with $\Gamma = 1$, the analysis is the same as for group randomized experiments in §2.3 or Small, Ten Have and Rosenbaum (2008).

2.5 Sensitivity analysis of the flooding in Bangladesh

As noted in §2.3, the covariance adjusted permutation test followed Small et al. (2008), setting q_{ski} equal to the rank of the residual of R_{ski} when regressed on the six covariates in $\mathbf{x}_{ski}$ using Huber's m-estimates (with the default settings of `rlm`, in R's `MASS` package [Venables and Ripley, 2002]). In a randomization test, $\Gamma = 1$, this yields a 1-sided P-value of 0.0064 testing Fisher's sharp null hypothesis H_0 of no treatment effect. The upper bound on this one-sided P-value is ≤ 0.045 for $\Gamma \leqslant 1.5$, so the finding that children in flooded villages were sicker is insensitive to small biases but is sensitive to moderately large biases.

If the null hypothesis of no effect is replaced by the hypothesis H_{τ_0} of a shift effect, $r_{Tski} = r_{Cski} + \tau_0$, then $R_{ski} - Z_{ski}\tau_0 = r_{Cski}$, so that, in the usual way, the randomization test yields a Hodges-Lehmann (1963) point estimate of effect, $\widehat{\tau}$; see Small et al. (2008). In the absence of bias, $\Gamma = 1$, the point estimate is $\widehat{\tau} = 1.04$ additional sick days. In a sensitivity analysis, there is not a single Hodges-Lehmann point estimate but an interval of estimates, the interval collapsing to a point when $\Gamma = 1$; see Rosenbaum (1993). When $\Gamma = 1.5$, the interval of point estimates is entirely positive, from 0.68 days to 1.41 days. The interval of point estimates just barely includes 0 days at $\Gamma = 4.1$.

How does clustered treatment assignment affect sensitivity to unmeasured biases? In designing a study, one might take one child per village, $n_{sk} = 1$, in effect yielding a study without grouped assignment. In the absence of bias, $\Gamma = 1$, such a design would be more efficient than a clustered study of the same size $N = \sum n_{sk}$, although it would entail collecting survey data at many more villages $2S$, and so might be prohibitively expensive. How do changes in the degree of clustering affect the conclusions of a sensitivity analysis (6) with $\Gamma > 1$? These questions are discussed beginning in §3. In light of this discussion, §3.5 performs some additional analyses of the flooding in Bangladesh.

3 Design Sensitivity with Clustered Treatment Assignment

3.1 Power of a sensitivity analysis; design sensitivity

The power of a sensitivity analysis is the probability, for a given value Γ of the sensitivity parameter and a given test size α, that the null hypothesis of no treatment effect H_0 will be rejected when it is in fact false and a treatment effect, not a bias, is responsible for the behavior of the test statistic, T. More precisely, for given α the power of a sensitivity analysis with parameter Γ is the probability that the upper bound on the P-value will be at most α when there is actually a treatment effect, so H_0 is false, and there is no bias from nonrandom treatment assignment, so $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F},\, \mathcal{Z}) = 1/|\mathcal{Z}| = 1/2^S$ for each $\mathbf{z} \in \mathcal{Z}$; that is, as $S \rightarrow \infty$, it is the probability of the event (6) under some specific model for a treatment effect without bias. For any stochastic model with a treatment effect, that is, for any model for the generation of $\mathcal{F}$, the probability of the event (6) with $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F},\, \mathcal{Z}) = 1/2^S$ may be determined analytically in simple situations or by simulation in complex situations. In general, refer to the situation in which H_0 is false and $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F},\, \mathcal{Z}) = 1/|\mathcal{Z}| = 1/2^S$ for each $\mathbf{z} \in \mathcal{Z}$ as the "favorable situation," so the power of a sensitivity analysis is computed in the favorable situation. Importantly, in an observational study we cannot recognize when we are in the favorable situation even as $S \rightarrow \infty$; that is, in an observational study, we cannot know that we are looking at a treatment effect without unmeasured bias rather than an unmeasured bias without a treatment effect. In general, the power of a sensitivity analysis depends upon the research design, that is the stochastic process that generated the data, and upon the selected methods of analysis. The power of a sensitivity analysis may guide the choice of research design for fixed methods of analysis, the choice of methods of analysis for a fixed research design, or the choice of research design when the method of analysis must change to accommodate the change in research design.

If we cannot know when we are in the favorable situation, and if we may not be in the favorable situation, then why should we be interested in the power computed in the favorable situation? In computing power in the favorable situation we are asking about the ability of a particular research design and method of analysis to discriminate between two situations in which we know unambiguously what answer is desired of the sensitivity analysis. If there is a moderate bias Γ in treatment assignment and no treatment effect, then we hope that the sensitivity analysis will tell us that the observed association between treatment and outcome can be explained by a bias of magnitude Γ, and by construction we take only a risk of at most α that the sensitivity analysis will report otherwise in this situation. If there is no bias in treatment assignment, $\Gamma = 1$, and there is a treatment effect then we hope to reject the null hypothesis H_0 of no effect, and the power of a sensitivity analysis in the favorable situation is the chance that our hope will be realized. If there were both a bias in treatment assignment and also a treatment effect, then we must be ambivalent about rejecting the hypothesis of no effect, H_0, even though it is false. Suppose, for example, that there was a large bias in treatment assignment and a small treatment effect, so that rejection of H_0 is nearly assured for all small or moderate Γ; then, we cannot be pleased to reject H_0 for small or moderate Γ because we know we would also have rejected H_0 in this situation had it been true.

In computing the power of a sensitivity analysis, we may, of course, substitute another definite null hypothesis about the effect, say the hypothesis H_{τ_0} of a shift effect, $r_{Tski} = r_{Cski} + \tau_0$, for the null hypothesis of H_0 of no effect. For instance, in the absence of bias in treatment assignment, $\Gamma = 1$, we may ask: what is the probability that the sensitivity analysis will reject H_{τ_0} allowing for bias $\Gamma \geq 1$ when H_{τ_0} is false and H_{τ_1} is true for a specific $\tau_1 > \tau_0$? However, this calculation reduces to the calculation already performed. If H_{τ_0} were true, then the $R_{ski} - Z_{ski}\tau_0 = r_{Cski}$ satisfy the null hypothesis of no effect, H_0, and if H_{τ_1} is true then $R_{ski} - Z_{ski}\tau_0$ satisfy the hypothesis $H_{\tau_1 - \tau_0}$. If the sensitivity analysis is applied to $R_{ski} - Z_{ski}\tau_0$, the the power to reject H_{τ_0} in favor of H_{τ_1} equals the power to reject H_0 for $R_{ski} - Z_{ski}\tau_0 = r_{Cski}$ against $H_{\tau_1 - \tau_0}$.

In general, the power depends upon S. For asymptotics, one considers a stochastic process that generates an $\mathcal{F}$ for each sample size S and then allows $S \rightarrow \infty$. For instance, the S cluster pairs s might be an independent and identically distributed sample of size S from an infinite population of cluster pairs. For each such stochastic process, we may study the probability of the event (6) with $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/2^S$ as $S \to \infty$.

Under mild conditions, as $S \to \infty$, there is a value $\tilde{\Gamma}$ called the design sensitivity such that the power of the sensitivity analysis – the probability of the event (6) with $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/2^S$ – tends to 1 if the sensitivity analysis is performed with $\Gamma < \tilde{\Gamma}$ and it tends to zero if $\Gamma > \tilde{\Gamma}$; see Rosenbaum (2004; 2010, Part III). In words, as the sample size increases, we can distinguish a specified treatment effect without bias from all biases smaller than $\tilde{\Gamma}$ but not from some biases larger than $\tilde{\Gamma}$. In general, the design sensitivity $\tilde{\Gamma}$ depends upon the stochastic process that generated $\mathcal{F}$ and on the choice of test statistic T. Among other things, the design sensitivity is a guide to designing observational studies to be less sensitive to unmeasured biases; see, for instance, Stuart and Hanna (2013) and Zubizarreta et al. (2013).

3.2 A formula for design sensitivity with clustered treatment assignment

If a clustered observational treatment assignment were not biased, so that $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\mathcal{Z}| = 1/2^S$ for each $\mathbf{z} \in \mathcal{Z}$, then we could not discern this from the data, and the best we could hope to say is that conclusions are insensitive to a moderately large bias Γ.

The current section calculates the design sensitivity $\tilde{\Gamma}$ in a simplified situation. Specifically, three conditions are required, and these are first stated, then discussed:

a1 We are in the favorable situation, so H_0 is false and $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/|\mathcal{Z}| = 1/2^S$ for each $\mathbf{z} \in \mathcal{Z}$.

a2 The pair of cluster sizes, (n_{s1}, n_{s2}), is constant, $(n_{s1}, n_{s2}) = (n_1, n_2)$ for all s, with $n_1 \geq 1$, $n_2 \geq 1$, and $w_s = 1$ for each s.

a3 The Q_s are independent and identically distributed with finite variance.

Condition a1 simply says we are in the situation in which the power of a sensitivity analysis and design sensitivity are computed. Condition a2 does not require $n_1 = n_2$; however, this equality would be common when cluster sizes are constant. If $n_1 = 1$ and $n_2 = 2$, then some cluster pairs contain 1 treated subject from one cluster and 2 controls from a paired cluster while other cluster pairs contain one control from one cluster and two treated subjects from a paired cluster. Condition a3 is a statement about the treated-minus-control difference in mean scores q_{ski} in cluster pair s, and it can be true in a variety of ways. For the permutational t-test with $q_{ski} = R_{ski}$, a3 would follow from a1 and a2 if $n_1 = n_2$ and cluster pairs were sampled at random from an infinite population of cluster pairs in which $\operatorname{var}(R_{ski}) < \infty$. For the permutational t-test with $q_{ski} = R_{ski}$ with $n_1 \neq n_2$, additional assumptions analogous to Gauss-Markov assumptions (i.e., additive effects, constant variance), would ensure that the mean differences Q_s satisfy a3. Condition a3 would also hold with $n_1 = n_2$ if the permutational t-test were replaced by the sum of S separately computed rank sum statistics, with q_{ski} equal to the rank of R_{ski} within cluster pair s, ranking from 1 to $n_1 + n_2$. Conditions a2 and a3 are one simple way of saying that as the number of clusters increases, $S \to \infty$, the added clusters are similar to the original ones, that the sequence of clusters is not evolving.

Assuming a3, let $\lambda = \mathrm{E}(Q_s)$ and $\eta = \mathrm{E}(|Q_s|)$, noticing that $\eta > \lambda$ unless $\Pr(Q_s < 0) = 0$. To have $\Pr(Q_s < 0) = 0$, the treatment effect would need to be so large that a cluster pair s with a negative sample mean difference, $Q_s < 0$, never occurs.

Proposition 1 *Assume a1-a3. If $\eta > \lambda$ then the design sensitivity is*

$$\tilde{\Gamma} = \frac{\eta + \lambda}{\eta - \lambda} \tag{7}$$

and otherwise $\tilde{\Gamma} = \infty$.

Proof. By the weak law of large numbers, as $S \to \infty$ in (6), the following quantities converge in probability:

$$\frac{T}{S} \to \lambda$$

$$\frac{\Gamma - 1}{S\left(\Gamma + 1\right)} \sum_{s=1}^{S} \left|Q_s\right| \to \frac{\left(\Gamma - 1\right)\eta}{\left(\Gamma + 1\right)}$$

$$\sqrt{\frac{4\Gamma}{S^2\left(1+\Gamma\right)^2} \sum_{s=1}^{S} Q_s^2} = \sqrt{\frac{1}{S}} \sqrt{\frac{4\Gamma}{\left(1+\Gamma\right)^2} \cdot \frac{1}{S} \sum_{s=1}^{S} Q_s^2} \to 0$$

It follows that the probability of the event (6) tends to 1 as $S \to \infty$ if $\lambda > \left(\Gamma - 1\right)\eta / \left(\Gamma + 1\right)$ and to 0 if $\lambda < \left(\Gamma - 1\right)\eta / \left(\Gamma + 1\right)$ from which the proposition follows. ∎

There are many ways to weaken assumptions a2 and a3 yet retain a conclusion similar to (7). Essentially, one needs the three in-probability limits that appear in the proof, where these limits now define λ and η, and $\overline{T}_\Gamma / \mathrm{var}^{1/2}(\overline{T}_\Gamma)$ must be approximable as a standard Normal random variable.

When conditions a1-a3 hold, the computation or simulation of the design sensitivity $\tilde{\Gamma}$ is straightforward as it is requires two expectations, $\lambda = \mathrm{E}\left(Q_s\right)$ and $\eta = \mathrm{E}\left(\left|Q_s\right|\right)$, both of which are determined by a conventional model for clustered data with a treatment effect and randomized assignment of one cluster in a pair to treatment. If the distribution of Q_s has an explicit mathematical form, then the needed expectations may be determined by numerical integration. When the distribution of Q_s does not have a tractable mathematical form, but data sets yielding values of Q_s may be simulated, sampling many Q_s and averaging Q_s and $|Q_s|$ yields estimates of λ and η and estimated standard errors of those estimates.

3.3 Some numerical evaluations of design sensitivity for the permutational t-test

A simple common model for pairs of clusters $\mathcal{F}$ has an additive treatment effect, $r_{Tski} = r_{Cski} + \tau$, and $r_{Cski} = \phi_s + \xi_{sk} + \epsilon_{ski}$ where the cluster errors ξ_{sk} are independent and identically distributed with a Normal distribution having expectation 0 and finite variance

σ_ξ^2, the individual errors ϵ_{ski} are independent and identically distributed with a Normal distribution having expectation 0 and finite variance σ_ϵ^2, and the ξ_{sk} and ϵ_{ski} are independent of each other. Then $R_{ski} = Z_{sk}\tau + \phi_s + \xi_{sk} + \epsilon_{ski}$. The intra-cluster correlation (ICC) $\zeta^2 = \sigma_\xi^2 / \left(\sigma_\xi^2 + \sigma_\epsilon^2\right)$ is the fraction of the variance in r_{Cski} that is due to the cluster error ξ_{sk} rather than the individual error ϵ_{ski}. If treatment assignment is not biased, so that $\Pr\left(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}\right) = 1/2^S$, and the permutational t-test is used, so that $q_{ski} = R_{ski}$ and ϕ_s cancels upon taking differences within cluster pair s, then $\tau = \lambda = \mathrm{E}\left(Q_s\right)$ and $\mathrm{var}\left(Q_s\right) = \sigma^2 = 2\sigma_\xi^2 + \left(1/n_1 + 1/n_2\right)\sigma_\epsilon^2$. A study without clusters (or a study that sampled one subject per cluster) would have $n_1 = n_2 = 1$ and $\mathrm{var}\left(Q_s\right) = 2\sigma_\xi^2 + 2\sigma_\epsilon^2$ and for numerical comparisons we set this equal to 1, so $\tau = \lambda = \mathrm{E}\left(Q_s\right)$ is the expected treatment effect in units of the standard deviation without clusters, $n_1 = n_2 = 1$, that is, the expected effect when a matched pair difference has variance 1. If $\zeta^2 > 0$ then increasing n_1 or n_2 leaves $\tau = \lambda = \mathrm{E}\left(Q_s\right)$ unchanged but has a less than proportional effect on $\mathrm{var}\left(Q_s\right) = \sigma^2 = 2\sigma_\xi^2 + \left(1/n_1 + 1/n_2\right)\sigma_\epsilon^2 = \zeta^2 + \left(1/n_1 + 1/n_2\right)\left(1 - \zeta^2\right)/2$ because the between cluster component $2\sigma_\xi^2 = \zeta^2$ is not reduced.

Table 2 concerns the permutational t-test, that is $q_{ski} = R_{ski}$, for S independent cluster pairs, each cluster being of same size $n = n_1 = n_2$, each pair having expected mean difference $\tau = \lambda = \mathrm{E}\left(Q_s\right)$, variance $\mathrm{var}\left(Q_s\right) = \sigma_n^2 = \zeta^2 + \left(1 - \zeta^2\right)/n$, with the added assumption that the Q_s are Normally distributed. Table 2 lets $S \to \infty$ in this situation and displays the design sensitivity, $\tilde{\Gamma}$. The value of $\eta = \mathrm{E}\left(|Q_s|\right)$ is obtained by numerical integration.

For instance, in Table 2, $\tilde{\Gamma} = 7.47$ for $\zeta^2 = .25$, $n = 5$, $\tau = 1/2$. This says that the power of a sensitivity analysis in this sampling situation tends to 1 as $S \to \infty$ if the analysis is performed with $\Gamma < \tilde{\Gamma} = 7.47$ and the power tends to 0 if $\Gamma > \tilde{\Gamma} = 7.47$. To illustrate this, drawing a single sample of $S = 100,000$ pairs from this sampling situation, the upper bound on the P-value using the permutational t-test is 0.014 for $\Gamma = 7.3$ and is 0.987 for $\Gamma = 7.7$.

In Table 2, the cluster size n does not matter if all of the variation is between clusters, $\zeta^2 = 1$, and the percent of variation between clusters ζ^2 does not matter if each cluster is of size $n = 1$. Of course, the design sensitivity is larger when the treatment effect τ is larger. The important pattern in Table 2 is that increasing the cluster size n when the variation between clusters is at most $50\% \geq \zeta^2$ substantially increases the design sensitivity $\tilde{\Gamma}$: the larger the cluster size, the larger the bias needed to explain away a treatment effect of fixed size τ. Table 3 is similar to Table 2, except n_1 and n_2 may differ. The pattern is similar.

Tables 2 and 3 indicate that a selection bias of magnitude Γ does more harm if it selects individuals than if it selects clusters, providing there is meaningful variation within clusters, say $50\% \geq \zeta^2$. For instance, if the clusters were schools, you could more severely bias a treatment-control comparison by picking the best individual students for treatment than if you could only pick schools with a disproportionate number of the best students. If the clusters were hospitals, you could more severely bias a treatment-control comparison by selecting the sickest patients for treatment than by selecting hospitals with many sick patients. Mechanisms of selection for treatment that merely favor stronger students or schools, rather than consciously engineering the strongest possible treatment group, are modeled by (4) with $1 < \Gamma < \infty$, and Tables 2 and 3 confirm that such probabilistic selection biases also can cause more harm when individuals rather than clusters are selected for treatment. In both cases, a larger departure from random assignment measured by Γ would need to be present to explain the same treatment effect if the assignment were at the cluster level.

As a general principle, it is known that if one can change the study design to reduce the heterogeneity of unit responses without altering the magnitude of the treatment effect, then one will make the study less sensitive to unmeasured biases; see Rosenbaum (2005). This general principle plays a role in clustered treatment assignments, because the units are now clusters rather than individuals. As noted above, under the model $R_{ski} = Z_{sk}\tau + \phi_s + \xi_{sk} + \epsilon_{ski}$ with $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}, \mathcal{Z}) = 1/2^S$, the expectation of the treated-minus-control difference in cluster means in pair s is $\tau = \lambda = \mathrm{E}(Q_s)$ with variance $\mathrm{var}(Q_s) = \sigma^2 = 2\sigma_\xi^2 + (1/n_1 + 1/n_2)\,\sigma_\epsilon^2$. It follows that increasing the cluster sizes, n_1 and n_2, under this model increases the size of the effect relative to the standard deviation, τ/σ, approaching an asymptote of $\tau/\left(\sqrt{2}\sigma_\xi\right)$ as $\min(n_1, n_2) \to \infty$. In brief, the difference in cluster means in pair s has expectation τ but is less heterogeneous than the difference of a pair of individual responses (with $n_1 = n_2 = 1$); hence, we expect the results to be less sensitive to unmeasured biases.

3.4 Some power comparisons: Does $\widetilde{\Gamma}$ provide useful guidance for moderate S?

Table 4 simulates power of a $\alpha = 0.05$ level, one-sided sensitivity analysis performed with $\Gamma = 4$, so it is estimating the probability of the event (6). The sampling situation is the same as in §3.3. In Table 4, there are S pairs of two clusters of equal size, $n_1 = n_2$, and the study contains $S(n_1 + n_2)$ individuals in total, either 500 or 1000 individuals. The situation with $n_1 = n_2 = 1$ is indistinguishable from a paired study without clusters. The intra-cluster correlation (ICC) is $\zeta^2 = 1/4$ or $\zeta^2 = 0$, but its value does not matter when $n_1 = n_2 = 1$. The treatment effect τ is expressed in units of the standard deviation of treated-minus-control pair difference when $n_1 = n_2 = 1$, so for two individuals from paired but different clusters, the expected effect is τ standard deviations. Each sampling situation is replicated 10,000 times, so the standard error of the estimated power is at most $0.005 = \sqrt{1/(4 \times 10,000)}$.

Table 4 considers both the permutational t-test with $q_{ski} = R_{ski}$ and Wilcoxon's two-sample ranks with $q_{ski} = \text{rank}\,(R_{ski})$ where the ranks are from 1 to $S\,(n_1 + n_2)$, as in Conover and Iman (1981) and Lam and Longnecker (1983). In the t-test, the cluster pair term ϕ_s cancels when differences are taken in (2), but this is no longer quite true when ranks are used. Rather than introduce an additional factor in the simulation for the rank statistic, we take $\phi_s = 0$ in this simulation. With short-tailed Gaussian data, the t-test and the rank test have similar powers in Table 4.

In Table 4, the number of pairs of clusters S is finite. For the t-test, the power should tend to 0 as $S \to \infty$ if $\widetilde{\Gamma} < \Gamma = 4$ in Table 2, and it should tend to 1 if $\widetilde{\Gamma} > \Gamma = 4$, and the patterns in Table 4 are consistent with that anticipation. For instance, with $\zeta^2 = 1/4$, $\tau = 1/2$, $n_1 = n_2 = 5$ in Table 2, $\widetilde{\Gamma} = 7.47 > 4 = \Gamma$, and the power in Table 4 increases with S.

There are two notable conclusions from Table 4. First, consistent with the asymptotic results in Table 2, a study with clustered treatment assignments may have substantial power in a sensitivity analysis when an otherwise identical study that sampled one person from each cluster would have negligible power. This is in marked contrast to randomized experiments where clustered treatment assignments tend to reduce power. The reduction in effective sample size from clustered assignment is reducing power in an observational sensitivity analysis, as in a randomization test, but in the sensitivity analysis this may be offset by an increase in design sensitivity. Second, although the design sensitivities $\widetilde{\Gamma}$ in Table 2 describe the situation as $S \to \infty$, Table 4 indicates that $\widetilde{\Gamma}$ provides useful guidance with $S = 50$ clusters.

3.5 Additional analyses of the flood in Bangladesh: using multiple weights; role of covariance adjustment

The current section presents some additional analyses of the flood in Bangladesh in light of considerations earlier in §3. Specifically, we reconsider the weights, w_s, in (2), and the role of covariance adjustment at the individual level, as discussed in §2.3. In §2.5, the weights were constant, $w_s \propto 1$, and the permutation inference was applied to residuals from a robust covariance adjustment as in Rosenbaum (2002a).

One commonly used and natural set of weights w_s is proportional to the total number of children in a cluster pair, $w_s \propto n_{s1}+n_{s2}$, or specifically, $w_s = (n_{s1}+n_{s2}) \,/\, \sum_{\ell=1}^{S}(n_{\ell 1}+n_{\ell 2})$. If treatments had been randomly assigned to clusters, $\Pr(\mathbf{Z}=\mathbf{z} \mid \mathcal{F},\, \mathcal{Z}) = 1/2^S$, then with $q_{ski} = R_{ski}$ and these weights, the statistic T in (2) would be unbiased for the average treatment effect, $\mathrm{E}(T) = \left\{\sum_{s=1}^{S}(n_{s1}+n_{s2})\right\}^{-1} \sum_{s=1}^{S}\sum_{k=1}^{2}\sum_{i=1}^{n_{sk}}(r_{Tski}-r_{Cski})$. Weights $w_s \propto n_{s1}+n_{s2}$ are particularly relevant when one suspects that the treatment effect may be larger in some cluster pairs than in others. In contrast, if one believed that the treatment effect was constant, $r_{Tski}-r_{Cski}=\tau$, then one would have some freedom to adjust the weights to reduce the variance of T as an estimate of τ, as will now be described in detail.

Sections 3.3 and 3.4 asked what would happen in a sensitivity analysis using the permutational t-test if, in fact, there were a treatment effect and no bias in treatment assignment, considering in particular the model $R_{ski} = Z_{sk}\tau + \phi_s + \xi_{sk} + \epsilon_{ski}$ with $\Pr(\mathbf{Z}=\mathbf{z} \mid \mathcal{F},\, \mathcal{Z}) = 1/2^S$, $\mathrm{var}(\xi_{sk}) = \sigma_\xi^2$, $\mathrm{var}(\epsilon_{ski}) = \sigma_\epsilon^2$ and independence of all ξ_{sk} and ϵ_{ski}. In this case, for the permutational t-test with $q_{ski} = R_{ski}$, we have $\mathrm{var}(Q_s) = 2\sigma_\xi^2 + (1/n_{s1}+1/n_{s2})\,\sigma_\epsilon^2$ and $\mathrm{E}(T) = \tau$, providing the weights, $w_s \geq 0$, sum to one, $1 = \sum_{s=1}^{S} w_s$. Among weights that sum to 1, the weights that minimize $\mathrm{var}(T)$ are inversely proportional to $\mathrm{var}(Q_s)$, that is, $w_s \propto \left\{2\sigma_\xi^2 + (1/n_{s1}+1/n_{s2})\,\sigma_\epsilon^2\right\}^{-1}$. If there were no variability among individuals in the same cluster, $\sigma_{\epsilon}^{2}=0$, then $\operatorname{var}(Q_s)$ would be minimized by constant weights, $w_s \propto 1$, whereas if there were no extra variability from clusters, $\sigma_{\xi}^{2}=0$, then $\operatorname{var}(Q_s)$ would be minimized by weights suggested in Kalton (1968, his expression (9)), $w_s \propto (1/n_{s1}+1/n_{s2})^{-1}=n_{s1}n_{s2}/(n_{s1}+n_{s2})$.

In brief, three possible weights with somewhat incompatible motivations are $w_s \propto n_{s1}+n_{s2}$, $w_s \propto 1$, and $w_s \propto n_{s1}n_{s2}/(n_{s1}+n_{s2})$. In testing the null hypothesis H_0 of no effect, each set of weights is valid, in the sense that if the bias is at most Γ then an α-level sensitivity analysis falsely rejects H_0 with probability at most α when the sensitivity analysis is performed at Γ. The weights will affect the power of the sensitivity analysis. In this context, one attractive approach is to perform three sensitivity analyses with different weights, to select the smallest or most significant of the three upper bounds on P-values, and to correct that smallest P-value for multiple testing, as discussed in Rosenbaum (2012). This method has the best or largest of the three design sensitivities of the three component tests. Moreover, because the three tests are very highly correlated, the correction for multiple testing is small, much smaller than a correction using the Bonferroni inequality; see Rosenbaum (2012, Table 4).

For the data from Bangladesh, at $\Gamma=1.5$: (i) constant weights $w_s \propto 1$ yield an upper bound on the P-value of 0.045, as in §2.5; (ii) Kalton's weights $w_s \propto n_{s1}n_{s2}/(n_{s1}+n_{s2})$ yield an upper bound of 0.0498; (iii) weighting proportional to the sample size in a cluster pair, $w_s \propto n_{s1}+n_{s2}$, yields an upper bound of 0.0432. In effect, the combined test corrects the smallest of these three P-values, namely 0.0432, for multiple testing, taking account of the high correlation among the three tests. Correction for multiple testing yields an upper bound of 0.0499 at $\Gamma=1.5$. In this example, the choice of weights did not matter much, perhaps because the clusters were of similar size.

A treatment effect could vary in magnitude with cluster size. For instance, a treatment might be more or less effective in large schools as opposed to small schools. If the cluster sizes varied markedly and the treatment effect did vary with cluster size, then the choice of weights might matter more than it did in the data from Bangladesh. In such a situation, the use of more than one set of weights, as above, may avoid a loss of power in a sensitivity analysis as a consequence of an unwise choice of weights.

As discussed in §2.3, the analysis in §2.5 used robust covariance adjustment to remove variation in the outcome, days ill, that could be predicted from covariates $\mathbf{x}_{ski}$ that describe individuals. In a randomized experiment, there is no bias in treatment assignment, $\Pr\left(\mathbf{Z}=\mathbf{z} \mid \mathcal{F},\ \mathcal{Z}\right)=1/2^S$, and covariance adjustment serves to reduce variability, roughly speaking to reduce σ_ϵ^2 and possibly σ_ξ^2. As discussed in §3.3 and in Rosenbaum (2005), a reduction in unit heterogeneity with no change in the treatment effect is expected to reduce sensitivity to unmeasured biases. This may have occurred to a small degree in the data from Bangladesh. With constant weights, $w_s \propto 1$, and with covariance adjustment as in §2.5, the upper bound on the P-value is 0.045 at $\Gamma = 1.5$, but without covariance adjustment, the upper bound on the P-value is 0.064 at $\Gamma = 1.5$, so in this example there is somewhat more sensitivity to unmeasured bias if covariance adjustment is not used. In observational studies, covariance adjustment aims to both reduce heterogeneity and to reduce bias from measured covariates, and these cannot be distinguished in an empirical study in which the actual effects and biases are unknown.

4 Contrasting clustered and individual analyses when clusters are assigned treatments

4.1 Overview: A finite sample inequality and an asymptotic comparison

What happens if a sensitivity analysis is performed at the individual level when treatments are assigned to clusters, not individuals? That is: How would the results of the sensitivity analysis be different if the clustering were simply ignored? Given (4), of course, one of these two sensitivity analyses is incorrect: if clusters are assigned to treatment or control, then (6) is correct, and ignoring the clustered treatment assignment is not correct. If the clustering is known, as in Table 1, the correct analysis (6) may be performed, and there is no need to perform an incorrect analysis at the individual level. There are, however, several reasons to be interested in the relationship between the clustered and individual analyses. At an entirely practical level, in some public use data sets, identifying information is removed to preserve confidentiality, with the consequence that it may not be possible to tell when two individuals attend the same school or were treated in the same hospital. That is, in some observational studies, the clustering itself is not observed. At a conceptual level, some modeling situations may not present a clear choice between clustered and individual models of treatment assignment. For example, to study a resource present at some clinics but not others, one ordinarily models the assignment of resources to clinic populations; yet in certain circumstances, as in Baiocchi et al. (2010), it may be possible to model more persuasively the individual patient's selection of a clinic. How in general does the choice between models of individual or of cluster-level assignment bear on sensitivity to hidden bias?

Here again, intuition derived from randomized experiments turns out to be an imperfect guide. For simplicity of discussion in the current paragraph, consider a study with clusters of constant size. Under random assignment, ignoring the clustering of treatment assignments does not bias treatment-control comparisons, but it does exaggerate effective sample size: the correct sample size is the number of clusters, not the total number of individuals within those clusters. So tests and confidence intervals that do not account for clustering are straightforwardly anticonservative. By contrast, with $\Gamma > 1$ in (4), treatment-control comparisons may be biased whether or not treatment is assigned by cluster, and a larger potential bias under the individual assignment model may overwhelm its advantage in standard errors. Indeed, (4) allows for bias in T/S that need not diminish as the sample size increases, while $\text{var}(T/S)$ decreases in the ordinary way, as $O(S^{-1})$. Does an allowance for bias of fixed size $\Gamma > 1$ partially address the tendency of clustering to understate the magnitude of sampling variability? It depends. When $\Gamma \approx 1$ and S is small, the answer is no; but when $\Gamma \gg 1$ and S is large, the answer is yes. Sections 4.2 and 4.3 make this precise in two different ways.

Two related results are presented. In §4.2, the two test statistics, with and without allowance for clustering, are compared as functions of the data. In particular, Proposition 2 says that the individual level analysis can be rendered conservative by making a correction that adjusts the effective sample size from the number of individuals to the smaller number of clusters. In §4.3, the two test statistics are compared as the number of cluster pairs increases, $S \to \infty$, with clusters of fixed size. Proposition 3 gives fairly general conditions such that, in very large samples S, the sensitivity analysis with $\Gamma > 1$ at the individual level is conservative even without correction for the sample size.

4.2 Comparing the test statistics as functions of the data: an inequality derived from convexity

Some further insight is provided by viewing the test statistic on the left in (6) as a function of the data and comparing it to the test statistic that would be used in a sensitivity analysis performed at the individual level. There is a sense in which an analysis at the individual level exaggerates the effective sample size, because only S independent assignments of whole clusters were made, but if a simple correction is made for the exaggeration of the sample size, then the individual analysis is conservative when compared to (6). The issue is made explicit in Proposition 2 below.

The structure is as follows. A total of S cluster pairs are sampled. After sampling a cluster pair, s, then $\overline{n}$ subjects from one cluster in the pair, $s1$, are individually matched for observed covariates $\mathbf{x}_{ski}$ to $\overline{n}$ distinct subjects in the other cluster in the pair, $s2$, so in the end we have a pair of clusters and $\overline{n}$ pairs of individuals, one individual in each pair coming from each cluster. It is notationally convenient to renumber the individuals so that individual $s1i$ is paired with $s2i$, $i = 1, \ldots, \overline{n}$. Because each cluster contributes the same number, $\overline{n}$, of individuals, take $w_s = 1$ for all s. Clustered treatment assignment means that there are not $2^{\overline{n}}$ possible treatment assignments within cluster pair s, but rather 2 possible treatment assignments, with all $\overline{n}$ pairs assigned at once based on the assignment of their clusters sk. In total, there are not $2^{S\overline{n}}$ but rather 2^S possible treatment assignments for S pairs, so the effective sample size is S cluster pairs, not $S\overline{n}$ individual pairs. If $q_{ski} = R_{ski}$ then T/S in (2) is simultaneously the mean of the S cluster pair differences and the mean of the $S\overline{n}$ individual pair differences. If q_{ski} is the rank of R_{ski} within the $2\overline{n}$ units in cluster pair s, then T in (2) is linearly related to the sum of S Wilcoxon rank sum statistics but it is also the sum of S individually paired Wilcoxon statistics as discussed by Lam and Longnecker (1983).

Fix a number $\kappa > 0$. If one performed the sensitivity analysis at the individual level ignoring the clustered assignment, then one would incorrectly conclude that the upper bound on the one-sided P-value testing the null hypothesis H_0 of no effect is less than α at a specific Γ if the following inequality held with $\kappa = 1$:

$$\frac{T/S - \left[(\Gamma-1) / \left\{S\,\overline{n}\,(\Gamma+1)\right\}\right] \sum_{s=1}^{S} \sum_{i=1}^{\overline{n}} |q_{s1i} - q_{s2i}|}{\sqrt{\left[4\Gamma / \left\{(S\,\overline{n})^2 (1+\Gamma)^2\right\}\right] \sum_{s=1}^{S} \sum_{i=1}^{\overline{n}} (q_{s1i} - q_{s2i})^2}} \geq \kappa\, \Phi^{-1}(1-\alpha)\,. \qquad (8)$$

With $\kappa = 1$, (8) makes two miscalculations that work in opposite directions: (i) it exaggerates the effects of a bias of magnitude Γ because it imagines that bias acts on individuals rather than clusters, (ii) it exaggerates the effective sample size from S pairs to $S\overline{n}$ pairs. Proposition 2 shows that if $\kappa = 1$ is replaced by $\kappa = \sqrt{\overline{n}}$, then the exaggeration of sample size in (ii) is eliminated and (8) becomes conservative when compared to the correct analysis based on (6).

Proposition 2 *With S pairs of two clusters of equal size $\overline{n}$, and hence with equal weights $w_s = 1$, the statistics on the left sides of (6) and (8) are related by:*

$$\frac{\frac{T}{S} - \frac{\Gamma-1}{S\,(\Gamma+1)} \sum_{s=1}^{S} |Q_s|}{\sqrt{\frac{4\Gamma}{S^2(1+\Gamma)^2} \sum_{s=1}^{S} Q_s^2}} \geq \frac{1}{\sqrt{\overline{n}}} \frac{\frac{T}{S} - \frac{\Gamma-1}{S\,\overline{n}(\Gamma+1)} \sum_{s=1}^{S} \sum_{i=1}^{\overline{n}} |q_{s1i} - q_{s2i}|}{\sqrt{\frac{4\Gamma}{(S\,\overline{n})^2(1+\Gamma)^2} \sum_{s=1}^{S} \sum_{i=1}^{\overline{n}} (q_{s1i} - q_{s2i})^2}}. \tag{9}$$

If, for each cluster pair s, $q_{s1i} - q_{s2i}$ is constant, not varying with i, then equality holds in (9).

Proof. Because $\sum_{i=1}^{\overline{n}} |a_i|$ is a convex function of $(a_1, \ldots, a_{\overline{n}})$ and $Q_s = \overline{n}^{-1} \sum_{i=1}^{\overline{n}} (q_{s1i} - q_{s2i})$, it follows that $|Q_s| = \overline{n}^{-1} \sum_{i=1}^{\overline{n}} |Q_s| \leq \overline{n}^{-1} \sum_{i=1}^{\overline{n}} |q_{s1i} - q_{s2i}|$, so that the numerators in (9) are related by

$$\frac{T}{S} - \frac{(\Gamma-1)}{S\,(\Gamma+1)} \sum_{s=1}^{S} |Q_s| \geq \frac{T}{S} - \frac{(\Gamma-1)}{S\overline{n}\,(\Gamma+1)} \sum_{s=1}^{S} \sum_{i=1}^{\overline{n}} |q_{s1i} - q_{s2i}| \,. \tag{10}$$

Turning to the denominators and applying the Cauchy-Schwartz inequality yields

$$Q_s^2 = \left\{ \overline{n}^{-1} \sum_{i=1}^{\overline{n}} (q_{s1i} - q_{s2i}) \right\}^2 \leq \overline{n}^{-1} \sum_{i=1}^{\overline{n}} (q_{s1i} - q_{s2i})^2 \,,$$

with equality if and only if $q_{s1i} - q_{s2i}$ is constant as i varies for fixed s. It follows that the denominator on the right in (9) is greater than or equal to the denominator on the left, with equality if and only if $q_{s1i} - q_{s2i}$ is constant, not varying with i. Together with (10), this proves (9). ■

4.3 Asymptotic comparison as the number S of cluster pairs increases

The situation is simpler in the limit as the number the number S of clusters increases, $S \rightarrow \infty$. Suppose that the S cluster pairs are independently sampled from an infinite population of cluster pairs, and let $S \rightarrow \infty$ with $\overline{n}$ fixed and $w_s = 1$. One cluster in each pair is assigned to treatment, the other to control, with independent assignments in distinct clusters, with possibly biased assignment probabilities that may not satisfy (4), and the treatment may or may not have an effect. In this population, assume that $Q_s - (\Gamma - 1)\,|Q_s|\,/\,(\Gamma + 1)$ and $Q_s - [(\Gamma - 1)\,/\,\{\overline{n}\,(\Gamma + 1)\}]\sum_{i=1}^{\overline{n}}\,|q_{s1i} - q_{s2i}|$ have finite expectations η_Γ and $\eta_\Gamma^{'}$, respectively, and finite variances. For $\Gamma = 1$, the expectations are equal, $\eta_\Gamma = \eta_\Gamma^{'}$. Let $\Pi_{\Gamma S}$ be the probability of the event (6), let $\Pi_{\Gamma S}^{'}$ be the probability of the event (8), and let $\Psi_{\Gamma S}$ be the probability of (8) but not (6), so $\Psi_{\Gamma S}$ is the probability that the individual analysis rejects H_0 for the given Γ and S but the clustered analysis does not reject. In essence, for $\Gamma > 1$ and for all $\kappa > 0$, Proposition 3 says that (8) may not be conservative for finite S but becomes nearly so as $S \rightarrow \infty$. Expressed informally, for sufficiently large S, part (ii) of Proposition 3 says that (6) is more likely than (8) to reject H_0 for whatever reason, while part (iii) speaks specifically about false rejection of a true null hypothesis when the sensitivity analysis model holds, saying the rate of false rejection is controlled.

Proposition 3 *Under the assumptions of the previous paragraph: (i)* $\eta_\Gamma \geq \eta_\Gamma^{'}$*; (ii) if* $\eta_\Gamma \neq \eta_\Gamma^{'}$ *then* $\Psi_{\Gamma S} \rightarrow 0$ *as* $S \rightarrow \infty$ *for all* $\kappa > 0$*; (iii) if* $\eta_\Gamma \neq \eta_\Gamma^{'}$ *and, in addition, the null hypothesis of no effect* H_0 *and the sensitivity model (4) are both true, then as* $S \rightarrow \infty$*,* $\limsup \Pi_{\Gamma S} \leq \alpha$ *and* $\limsup \Pi_{\Gamma S}^{'} \leq \alpha$*.*

Proof. The left and right sides of (10) are each means of S independent and identically distributed observations with expectations η_Γ and $\eta_\Gamma^{'}$, respectively, so (i) follows from (10) in the case of $S = 1$. Given (i), if $\eta_\Gamma \neq \eta_\Gamma^{'}$, then $\eta_\Gamma > \eta_\Gamma^{'}$, and this in turn implies that either $\eta_\Gamma > 0$ or $\eta_\Gamma^{'} < 0$. As $S \to \infty$, by the weak law of large numbers, the left and right sides of (10) converge in probability to η_Γ and $\eta_\Gamma^{'}$, respectively. At the same time, the denominators on both sides of (9) tend to 0 as $S \to \infty$. If $\eta_\Gamma > 0$, the probability of the event (6) will tend to 1 as $S \to \infty$, whereas if $\eta_\Gamma^{'} < 0$ the probability of (8) will tend to 0 for all $\kappa > 0$. Therefore, if $\eta_\Gamma \neq \eta_\Gamma^{'}$, as $S \to \infty$, the probability that (8) occurs but (6) does not is tending to zero for all $\kappa > 0$, proving (ii). If H_0 and the sensitivity model (4) are both true, then $\limsup \Pi_{\Gamma S} \leq \alpha$ from (5) and the central limit theorem approximation (6) to $\overline{T}_\Gamma$. Combining $\limsup \Pi_{\Gamma S} \leq \alpha$ with $\Psi_{\Gamma S} \to 0$ from (ii) yields $\limsup \Pi_{\Gamma S}^{'} \leq \alpha$. ■

The caveat $\eta_\Gamma \neq \eta_\Gamma^{'}$ in (ii) of Proposition 3 is not a trivial matter. It precludes two important cases: (I) a conventional randomization test with $\Gamma = 1$, and (II) clusters with unit intracluster correlation, as seen from the case of equality in Proposition 2. In neither case (I) nor case (II) is (8) conservative even as $S \to \infty$, and (8) is not conservative for small S. That said, in a sensitivity analysis performed with $\Gamma > 1$, with clusters that are internally heterogeneous, the individual analysis ignoring clustering (8) is conservative even for $\kappa = 1$ for sufficiently large S, and it can be made conservative for all S by taking $\kappa = \sqrt{\overline{n}}$.

The results in Propositions 2 and 3 suggest that, with many clusters of moderate size, a sensitivity analysis that allows for a nontrivial degree of bias, $\Gamma \gg 1$, may be conservative even if clustering is ignored. In this same situation, a randomization test, $\Gamma = 1$, may easily be anti-conservative, rejecting H_0 too often, because $\eta_\Gamma = \eta_\Gamma^{'}$ in Proposition 3.

5 Discussion

Intuitions forged in randomized experiments do not always carry over to nonrandomized observational studies. In a flawless randomized experiment, all uncertainty comes from a limited sample size; that is, a consistent estimate of the treatment effect is available. In a nonrandomized observational study, there are at least two sources of uncertainty, namely a limited sample size and unmeasured biases in treatment assignment whose effects do not diminish with increasing sample size. Even as the sample size increases, $S \to \infty$, in an observational study, a consistent estimate of the treatment effect is not available so long as biased treatment assignment of fixed size $\Gamma > 1$ remains a possibility. In a randomized experiment, clustered treatment assignment may be necessary for practical reasons, but it reduces power, efficiency, and effective sample size relative to treatment assignment at the individual level. In an observational study, clustered treatment assignment has two consequences pulling in opposite directions. As in experiments, with clustered treatment assignment there is a reduction in effective sample size. Unlike randomized experiments, deviations from random treatment assignment of a given magnitude Γ have a smaller impact when forced to select whole clusters than when permitted to select individuals. Clustered treatment assignments have been found to be less sensitive to bias using an asymptotic measure (§3.2-§3.3), using simulation in finite samples (§3.4), and using an inequality that compares the values of clustered and unclustered test statistics (§4).

A word of caution is in order. We have contrasted sensitivity analyses for treatment assignment at the individual or group level in situations that are essentially the same apart from the differing modes of treatment assignment. It may happen, of course, that treatments are assigned at the individual level in one situation and at the group level in some very different situation. In that case, the differing situations need to be taken into account in thinking about the best research design. For instance, in the US, alcohol consumption is largely self-inflicted by individual adults, whereas in some nations, alcohol consumption is banned by the government for religious reasons. One might study the health benefits or harms of alcohol consumption in the US with individual treatment assignment or switch to study it internationally with elements of grouped assignment, but clearly this switch changes the situation in several important ways, not just in terms of individual or grouped assignment. Our abstract results are relevant to thinking about one aspect of such a switch, but the results provide no guidance about many other aspects.

References

Baiocchi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010), “Building a stronger instrument in an observational study of perinatal care for premature infants,” *Journal of the American Statistical Association*, 105, 1285–1296.

Bruce, M. L., Ten Have, T. R., Reynolds, C. F. III, Katz, I. I., Schulberg, H. C., Mulsant, B. H., Brown, G. K., McAvay, G. J., Pearson, J. L., and Alexopoulos, G. S. (2004), “Reducing Suicidal Ideation and Depressive Symptoms in Depressed Older Primary Care Patients: A Randomized Trial,” *Journal of the American Medical Association*, 291, 1081-1091.

Conover, W. J. and Iman, R. L. (1981), “Rank transformations as a bridge between parametric and nonparametric statistics,” *The American Statistician*, 35, 124-129.

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

Cornfield, J. (1978), “Randomization by group,” *American Journal of Epidemiology*, 208, 100-102.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M., Wynder, E. (1959), “Smoking and lung cancer,” *Journal of the National Cancer Institute*, 22, 173-203.

Cox, D. R. (1952), *Planning of Experiments*, New York: Wiley.

Del Ninno, C., Dorosh, P. A., Smith, L. C., and Roy, D. K. (2001), *The 1998 floods in Bangladesh: disaster impacts, household coping strategies and response*, Research Report 122, Washington, DC: International Food Policy Research Institute.

van Elteren, P. H. (1960), "On the combination of independent two-sample tests of Wilcoxon," *Bulletin of the International Statistical Institute*, 33, 229-241.

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

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

Hansen, B. B. (2007), "Optmatch: flexible, optimal matching for observational studies," *R News*, 7, 18-24.

Hansen, B. B. and Bowers, J. (2009), "Attributing effects to a cluster randomized get-out-the-vote campaign," *Journal of the American Statistical Association*, 104, 873-875.

Hodges, J. L. and Lehmann, E. L. (1963), "Estimates of location based on ranks," *Annals of Mathematical Statistics*, 34, 598-611.

Hosman, C. A., Hansen, B. B., and Holland, P. W. H. (2010), "The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder," *Annals of Applied Statistics*, 4, 849-870.

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

Kalton, G. (1968), "Standardization: a technique to control for extraneous variables," *Journal of the Royal Statistical Society* C (Applied Statistics), 17, 118-136.

Lam, F. C. and Longnecker, M. T. (1983), "A modified Wilcoxon rank sum test for paired data," *Biometrika*, 70, 310-313.

Lehmann, E. L. (1975), *Nonparametrics*, San Francisco: Holden-Day.

Mantel, N. (1977), "Mantel-Haenszel analyses of litter-matched time-to-response data, with modifications for recovery of interlitter information," *Cancer Research*, 37, 3863-3868.

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.

Maritz, J. S. (1979), "A note on exact robust confidence intervals for location," *Biometrika*, 66, 163-166.

Murray, D. (1998), *Design and Analysis of Group Randomized Trials*, New York: Oxford University Press.

Neyman, J. (1923, 1990), "On the application of probability theory to agricultural experiments," *Statistical Science*, 5, 463-480.

Pitman, E. J. (1937), "Statistical tests applicable to samples from any population," *Journal of the Royal Statistical Society*, Supplement 4, 119-130.

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*, Series B, 45, 212-218.

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

Rosenbaum, P. R. (1993), "Hodges-Lehmann point estimates of treatment effect in observational studies," *Journal of the American Statistical Association*, 88, 1250-1253.

Rosenbaum, P. R. (2002a), "Covariance adjustment in randomized experiments and observational studies (with Discussion)," *Statistical Science*, 17, 286-327.

Rosenbaum, P. R. (2002b), *Observational Studies* (2^{nd} Edition), New York: Springer.

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 Statistian*, 59, 147-152.

Rosenbaum, P. R. and Silber, J. H. (2009), "Amplification of sensitivity analysis in observational studies," *Journal of the American Statistical Association*, 104, 1398-1405.

Rosenbaum, P. R. (2010), *Design of Observational Studies*, New York: Springer.

Rosenbaum, P. R. (2012), "Testing one hypothesis twice in observational studies," *Biometrika*, 99, 763-774.

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

Small, D. (2007), "Sensitivity analysis for instrumental variables regression with overidentifying restrictions," *Journal of the American Statistical Association*, 102, 1049-1058.

Small, D., Ten Have, T., and Rosenbaum, P. R. (2008), "Randomization inference in a group-randomized trial of treatments for depression: covariate adjustment, noncompliance and quantile effects," *Journal of the American Statistical Association*, 103, 271-279.

Stuart, E. A. (2010), "Matching methods for causal inference," *Statistical Science*, 25, 1-21.

Stuart, E.A. and Hanna, D. B. (2013), "Should epidemiologists be more sensitive to design sensitivity?" *Epidemiology*, 24, 88-89.

Venables, W. N. and Ripley, B. D. (2002), *Modern Applied Statistics with S*, New York: Springer.

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

Wolfe, D. A. (1974), "A characterization of population weighted symmetry and related results," *Journal of the American Statistical Association*, 69, 819-822.

Yu, B. B., Gastwirth, J. L. (2005), "Sensitivity analysis for trend tests: application to the risk of radiation exposure," *Biostatistics*, 6, 201-209.

Zubizarreta, J. R., Cerdá, M. and Rosenbaum, P. R. (2013), " Effect of the 2010 Chilean earthquake on posttraumatic stress: Reducing sensitivity to unmeasured bias through study design," *Epidemiology*, 24, 79-87.

Table 1: Days ill for sampled children during two-weeks following a flood in $S = 27$ pairs of villages, one severely flooded, $Z_{sk} = 1$, the other not flooded, $Z_{sk} = 0$. The ranks q_{ski} are ordinary ranks of residuals of individual sick days R_{ski} when regressed using m-estimation on covariates describing individuals and villages.

Pair	Sample size		Mean days sick		Mean rank		Rank
	n_{sk}		$n_{sk}^{-1} \sum_i R_{ski}$		$n_{sk}^{-1} \sum_i q_{ski}$		difference
	Flooded	Not	Flooded	Not	Flooded	Not	
s	$Z_{sk} = 1$	$Z_{sk} = 0$	$Z_{sk} = 1$	$Z_{sk} = 0$	$Z_{sk} = 1$	$Z_{sk} = 0$	Q_s
1	5	6	0.0	3.0	93.4	141.7	-48.3
2	6	5	3.5	1.4	186.0	170.2	15.8
3	4	7	6.2	2.0	211.0	97.0	114.0
4	4	4	5.5	5.0	214.5	207.8	6.8
5	5	4	4.6	2.2	167.8	134.0	33.8
6	4	6	4.0	0.2	152.5	79.5	73.0
7	4	6	7.0	3.5	194.8	177.3	17.4
8	5	6	7.0	4.8	195.6	132.8	62.8
9	4	12	3.0	1.1	183.0	120.2	62.8
10	5	6	0.0	1.7	121.2	160.2	-39.0
11	6	6	8.8	0.3	236.2	104.7	131.5
12	5	5	11.8	0.0	274.4	124.2	150.2
13	4	4	3.5	1.2	211.8	164.0	47.8
14	5	5	4.2	1.0	180.0	126.2	53.8
15	9	5	7.9	3.4	220.7	128.6	92.1
16	7	5	0.1	0.0	86.1	76.4	9.7
17	5	2	7.6	7.0	143.2	131.0	12.2
18	6	6	1.7	8.2	136.3	230.0	-93.7
19	6	7	2.7	0.6	161.7	108.3	53.4
20	6	5	5.5	2.8	194.8	184.6	10.2
21	5	3	0.0	0.0	90.8	91.0	-0.2
22	6	10	7.7	2.8	208.5	123.3	85.2
23	4	9	7.0	0.3	136.0	35.4	100.6
24	5	5	1.4	3.0	116.0	124.4	-8.4
25	5	6	0.0	4.5	85.4	174.2	-88.8
26	4	4	3.5	0.0	151.8	125.0	26.8
27	5	3	0.0	0.0	76.2	83.3	-7.1

	Medians, Quartiles, Extremes						
	Sample size		Mean days sick		Mean rank		Difference
	Flooded	Not	Flooded	Not	Flooded	Not	
Min	4	2	0.0	0.0	76.2	35.4	-93.7
Q-1	4	4	1.6	0.3	128.6	106.5	3.3
Med	5	5	4.0	1.7	167.8	126.2	26.8
Q-3	6	6	7.0	3.2	202.1	162.1	67.9
Max	9	12	11.8	8.2	274.4	230.0	150.2

Table 2: Design sensitivity $\widetilde{\Gamma}$ of the permutational t-test with Gaussian errors, paired clusters of equal size $\bar{n} = n_1 = n_2$, and intracluster correlation coefficient (ICC) of ζ^2.

	Treatment effect $\tau = 1/4$				
Cluster size	ICC ζ^2				
$\bar{n}$	100%	50%	25%	10%	0%
1	1.87	1.87	1.87	1.87	1.87
2	1.87	2.06	2.21	2.33	2.43
5	1.87	2.25	2.70	3.29	4.10
10	1.87	2.33	3.02	4.26	7.47
25	1.87	2.39	3.29	5.57	25.71
	Treatment effect $\tau = 1/2$				
Cluster size	ICC ζ^2				
$\bar{n}$	100%	50%	25%	10%	0%
1	3.53	3.53	3.53	3.53	3.53
2	3.53	4.30	4.95	5.52	6.01
5	3.53	5.12	7.47	11.22	17.89
10	3.53	5.52	9.37	19.36	66.08
25	3.53	5.80	11.22	34.57	1248.42
	Treatment effect $\tau = 3/4$				
Cluster size	ICC ζ^2				
$\bar{n}$	100%	50%	25%	10%	0%
1	6.72	6.72	6.72	6.72	6.72
2	6.72	9.11	11.34	13.40	15.31
5	6.72	11.94	21.53	41.17	87.73
10	6.72	13.40	30.87	99.95	801.84
25	6.72	14.48	41.17	263.30	178310.41

Table 3: Design sensitivity $\widetilde{\Gamma}$ of the permutational t-test with Gaussian errors, paired clusters of possibly unequal cluster sizes $n_1 \leq n_2$, intracluster correlation coefficient (ICC) of ζ^2, and treatment effect $\tau = 1/2$.

	Cluster size $n_1 = 1$				
Cluster size	ICC ζ^2				
n_2	100%	50%	25%	10%	0%
1	3.53	3.53	3.53	3.53	3.53
2	3.53	3.85	4.06	4.20	4.30
5	3.53	4.10	4.53	4.86	5.12
10	3.53	4.20	4.73	5.16	5.52
25	3.53	4.26	4.86	5.37	5.80
	Cluster size $n_1 = 2$				
Cluster size	ICC ζ^2				
n_2	100%	50%	25%	10%	0%
2	3.53	4.30	4.95	5.52	6.01
5	3.53	4.66	5.88	7.19	8.62
10	3.53	4.80	6.30	8.11	10.31
25	3.53	4.89	6.60	8.82	11.75
	Cluster size $n_1 = 5$				
Cluster size	ICC ζ^2				
n_2	100%	50%	25%	10%	0%
5	3.53	5.12	7.47	11.22	17.89
10	3.53	5.31	8.29	14.15	28.82
25	3.53	5.43	8.90	16.86	44.34

Table 4: Power of a 0.05-level, one-sided sensitivity analysis at $\Gamma = 4$ when one of two clusters in each pair of clusters is picked for treatment. Each situation is sampled 10,000 times.

Individuals	Cluster Pairs	Cluster Size	ICC	Effect		Power	
500	S	$n_1 = n_2$	ζ^2	τ	Γ	t-test	Wilcoxon
500	50	5	1/4	1/2	4	0.2490	0.2262
500	50	5	0	1/2	4	0.8862	0.8572
500	250	1	NA	1/2	4	0.0040	0.0017
1000	100	5	1/4	1/2	4	0.5266	0.4850
1000	100	5	0	1/2	4	0.9979	0.9959
1000	500	1	NA	1/2	4	0.0023	0.0009