2008

Randomization Inference in a Group–Randomized Trial of Treatments for Depression: Covariate Adjustment, Noncompliance and Quantile Effects

Dylan Small
University of Pennsylvania

Thomas R. Ten Have
University of Pennsylvania

Paul R. Rosenbaum
University of Pennsylvania

Randomization Inference in a Group–Randomized Trial of Treatments for Depression: Covariate Adjustment, Noncompliance and Quantile Effects

Abstract

In the Prospect Study, in ten pairs of two primary-care practices, one practice was picked at random to receive a "depression care manager" to treat its depressed patients. Randomization inference, properly performed, reflects the assignment of practices, not patients, to treatment or control. Yet, pertinent data describe individual patients: depression outcomes, baseline covariates, compliance with treatment. The methods discussed use only (i) the random assignment of clusters to treatment or control and (ii) the hypothesis about effects being tested or inverted for confidence intervals, so they are randomization inferences in Fisher's strict sense. There is no assumption that the covariance model generated the data, that compliers resemble noncompliers, that dependence is from additive random cluster effects, that individuals in a same cluster do not interfere with one another, or that units are sampled from a population. We contrast methods of covariance adjustment, never assuming the models are "true," obtaining exact randomization inferences. We consider exact inference about effects proportional to doses with noncompliance and effects whose magnitude varies with the degree of improvement that would occur without treatment. A simulation examines power.

Keywords
casual effect, instrumental variable, noncompliance, randomization

Disciplines
Statistics and Probability

RANDOMIZATION INFERENCE IN A GROUP-RANDOMIZED TRIAL OF TREATMENTS FOR DEPRESSION: COVARIATE ADJUSTMENT, NONCOMPLIANCE AND QUANTILE EFFECTS

DYLAN SMALL, TOM R. TEN HAVE, AND PAUL R. ROSENBAUM

Abstract: In the Prospect Study, in ten pairs of two primary care practices, one practice was picked at random to receive a 'depression care manager' to treat its depressed patients. Randomization inference, properly performed, reflects the assignment of practices, not patients, to treatment or control. Yet, pertinent data describe individual patients: depression outcomes, baseline covariates, compliance with treatment. The methods discussed use only (i) the random assignment of clusters to treatment or control and (ii) the hypothesis about effects being tested or inverted for confidence intervals, so they are randomization inferences in Fisher's strict sense. There is no assumption that the covariance model generated the data, nor that compliers resemble noncompliers, nor that dependence is from additive random cluster effects, nor that individuals in a same cluster do not interfere with one another, nor that units are sampled from a population. We contrast methods of covariance adjustment, never assuming the models are 'true,' obtaining exact, randomization inferences. We consider exact inference about effects proportional to doses with noncompliance, and effects whose magnitude varies with the degree of improvement that would occur without treatment. A simulation examines power.

1

1. Introduction: Example; Goals; Notation; Tests of No Effect

1.1. **A Group-Randomized Trial of Treatment for Depression.** The Prospect Study (Prevention of Suicide in Primary Care Elderly: Collaborative Trial) was a group-randomized trial of treatments for depression among adults over the age of 60 (Bruce, et. al 2004). Twenty primary care practices of varied sizes were paired into ten pairs on the basis of region (urban/other), affiliation, size and population type, and randomization was used to select one practice in each pair for the intervention, the other practice serving as a control. The intervention provided the practice with a depression care manager who was a social worker, a nurse or a psychologist, with weekly supervision of the manager by a psychiatrist. The depression care managers provided guideline-based treatment recommendations to physicians at the practice and interpersonal psychotherapy to some patients; see Bruce, et al. (2004, p. 1082) for a detailed description of the managers' activities. The control practices provided "usual care," with certain enhancements involving diagnosis and education of physicians about treatment guidelines.

In a group randomized trial, clusters are assigned to treatment or control, but individuals are of interest. Group randomized trials are discussed by Cornfield (1978), Gail, et al. (1996), Brookmeyer and Chen (1998), Donner (1998), Murray (1998), Braun and Feng (2001), Frangakis, Rubin and Zhou (2002), and Murray et al. (2006), among others.

Figure 1 displays the Hamilton Depression Score, a 24-item measure of depression severity, for the 253 treated patients and 234 control patients who had scores at baseline, before treatment, and at four months. For these patients, the Hamilton Scores look similar in treated and control groups at baseline, and there appears to be

some improvement at 4 months in both groups, but the improvement looks somewhat larger in the treated group. We also consider two baseline covariates, namely age and a binary indicator of suicidal ideation. For the patients in Figure 1, the median age was 70 in the treated group and 69 in the control group, and the frequency of suicidal ideation was 30% in the treated group and 21% in the control group.

For the treated and control practice in each pair, Table 1 displays the number of depressed patients and the mean change in the Hamilton Score from baseline to four months. In the first pair, the control practice had 44 depressed patients with a mean change of -4.7, while the treated practice had 49 depressed patients with a mean change of -4.6, so there is little difference observed in this pair. The results are fairly erratic, but larger improvements are often found in treated practices. Although all depressed patients at treated practices were referred to the depression care manager, some refused treatment from the depression care manager. Table 1 gives the percent compliance at four months in the treated practice, that is, the percent of depressed patients who accepted treatment from the depression care manager.

A test of a hypothesized treatment effect is a strict randomization inference if it uses just the null hypothesis under test and the randomization actually used in the experiment; it does not entail an assumption about a stochastic process that generated the data (Fisher 1935, Pitman 1937, Welch 1937, Lehmann 1998, §1, Cox and Reid 2000, §2.2.5). If the test rejects when this significance level is less than or equal to α, then randomization ensures that the test has level α, that is, the probability of falsely rejecting a true hypothesis is at most α. Let $\mathcal{H}$ be a set of hypotheses about the treatment effect, where at most one hypothesis in $\mathcal{H}$ is true. Applying a strict randomization test to each hypothesis in $\mathcal{H}$ divides $\mathcal{H}$ into

hypotheses $\mathcal{H}_r$ that are rejected at level α and hypotheses $\mathcal{H}_a$ not rejected at level α, forming a $1 - \alpha$ confidence set. Randomization ensures that a true hypothesis in $\mathcal{H}$ will fall in the rejected subset $\mathcal{H}_r$ with probability at most α. (The previous sentence is true whether or not $\mathcal{H}$ contains a true hypothesis, so the hypotheses $\mathcal{H}_r$ excluded from the confidence set have been validly rejected even if, unknown to us, the entire model, $\mathcal{H}$, is false in the sense that every $H \in \mathcal{H}$ is false; see Rosenbaum 2002, Rejoinder §6, pp. 324-325, for discussion.)

1.2. **Notation: Random Assignment of Clusters of Patients.** There are S strata, $s = 1, \ldots, S$, with K_s clusters in stratum s, where cluster k in stratum s contains n_{sk} individuals, $i = 1, \ldots, n_{sk}$, $k = 1, \ldots, K_s$. There are $n_s = \sum_{k=1}^{K_s} n_{sk}$ individuals in stratum s and $N = \sum_{s=1}^{S} n_s$ individuals in total. In stratum s, a fixed number, m_s, of clusters are picked at random for treatment, the others receiving control, with independent selections in distinct strata. If every cluster sk contained just one individual, $n_{sk} = 1$, then this would describe an experiment that is completely randomized within each stratum. In the Prospect Study, there were $S = 10$ pairs of $K_s = 2$ clinics, and $m_s = 1$ clinic was picked at random for treatment, and the n_{sk}'s are displayed in Table 1. If the kth cluster in stratum s is randomly assigned to treatment, write $Z_{ski} = 1$ for all $i = 1, \ldots, n_{sk}$; otherwise, if this cluster is assigned to control write $Z_{ski} = 0$ for all $i = 1, \ldots, n_{sk}$. Let $\mathbf{Z} = (Z_{111}, Z_{112}, \ldots, Z_{S,K_S,n_S})^T$.

The response is the decline in Hamilton Score from baseline to four months. The ith patient in the kth cluster of stratum s (or patient ski) has two potential responses, r_{Tski} if the cluster containing i is assigned to treatment, and r_{Cski} if this cluster is assigned to control. The effect on patient ski of assigning the kth cluster in stratum s to treatment rather than control is a comparison of r_{Tski} and r_{Cski}, such

as $r_{Tski} - r_{Cski}$; see Neyman (1923), Rubin (1974). Because each cluster receives either treatment or control, either r_{Tski} or r_{Cski} is observed, not both, so $r_{Tski} - r_{Cski}$ cannot be calculated from observable data. As r_{Tski} is observed only if $Z_{ski} = 1$ and r_{Cski} is observed only if $Z_{ski} = 0$, the observed response is $R_{ski} = Z_{ski} r_{Tski} + (1 - Z_{ski}) r_{Cski}$. Each individual has a (row) vector $\mathbf{x}_{ski}$ of pretreatment covariates; in the Prospect Study, $\mathbf{x}_{ski}$ contains patient's age and a binary indicator of suicidal ideation at baseline. Write $\mathbf{r}_T$, $\mathbf{r}_C$, and $\mathbf{R}$, for the corresponding N-dimensional vectors, and $\mathbf{X}$ for the matrix whose N rows are the $\mathbf{x}_{ski}$.

There are also two potential compliance outcomes, one under treatment, d_{Tski}, the other under control, d_{Cski}, so the observed compliance is $D_{ski} = Z_{ski} d_{Tski} + (1 - Z_{ski}) d_{Cski}$. Write $\mathbf{d}_T$, $\mathbf{d}_C$ and $\mathbf{D}$ for the corresponding N-dimensional vectors. In the Prospect Study, compliance is binary, 1 if the patient accepts treatment from the depression care manager, 0 otherwise. Although treatment assignment Z_{ski} is randomized, compliance D_{ski} may be highly nonrandom, and a comparison of compliers and noncompliers, or compliers and controls, may be severely biased as an estimate of the treatment effect. For instance, perhaps a patient ski who would decline treatment by a depression care manager if it were available, $d_{Tski} = 0$, is a patient who, based on the patient's own past experience, tends to recover somewhat from depression without treatment, that is, has a somewhat more favorable, more negative r_{Cski}. Alternatively, perhaps a patient who declines treatment when available tends to greater hopelessness, inactivity and a poorer prognosis. A correct analysis must not mistake self selection biases for treatment effects. Write $\mathcal{F} = \langle \mathbf{r}_T, \mathbf{r}_C, \mathbf{d}_T, \mathbf{d}_C, \mathbf{X} \rangle$, so $\mathcal{F}$ does not change as the treatment assignments, $\mathbf{Z}$, change, whereas $\mathbf{R}$ and $\mathbf{D}$ are functions of $\mathbf{Z}$ and $\mathcal{F}$, so they may change with $\mathbf{Z}$.

A special feature of the Prospect Study is that if a patient ski is at a clinic which receives a depression care manager, so $Z_{ski} = 1$, then the patient may or may not accept treatment, and d_{Tski} may equal 0 or 1, but if patient ski is at a clinic that does not receive a depression care manager, $Z_{ski} = 0$, then necessarily the patient does not receive care from a depression care manager, so necessarily $d_{Cski} = 0$. Unlike the Prospect Study, in studies of widely available treatments, such as aspirin or vitamins, controls may fail to comply by taking the treatment anyway, so that some $d_{Cski} \neq 0$. The methods we discuss may be used with or without the special feature of the Prospect Study, and with binary or continuous measures of compliance.

There are $L = \prod_{s=1}^{S} \binom{K_s}{m_s}$ possible values $\mathbf{z}$ of the treatment assignment $\mathbf{Z}$; collect these possible $\mathbf{z}$ in a set Ω. For a finite set G, write $|G|$ for number of elements of G, so $|\Omega| = L$. Because random numbers are used to assign clusters to treatments, $\Pr(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}) = 1/L$ for each $\mathbf{z} \in \Omega$. In the Prospect Study, there were $N = 487$ patients in total in these clinics, so $\mathbf{Z}$ is of dimension $N = 487$, with all patients in a clinic assigned to the same treatment. Hence, there were $L = \prod_{s=1}^{10} \binom{2}{1} = 2^{10} = 1024$ possible treatment assignments $\mathbf{z}$ in Ω and each had probability $1/1024$.

In a subtle way, the issue of 'interference between patients in the same cluster' does not arise. Cox (1958a, §2.4) said there is no interference between units if "the observation on one unit [is] unaffected by the particular assignment of treatments to other units;" Rubin (1986) calls this the 'stable unit-treatment value assumption.' A 'unit' is an opportunity to apply the treatment, so it is a cluster, not an individual. In the Prospect Study, it is unlikely that patients in different practices interfere with each other, and if so, then patient ski has only two potential responses in this experiment, r_{Tski} if cluster sk is assigned to treatment, and r_{Cski} if cluster sk is

assigned to control. Patient ski might have other responses in another experiment in which treatments were assigned at the individual level, but this possibility does not invalidate hypotheses about $(r_{Tski},\ r_{Cski})$.

1.3. **Tests of No Effect.** Let $\mathbf{e}$ be a function of $\mathcal{F} = \langle \mathbf{r}_T,\ \mathbf{r}_C,\ \mathbf{d}_T,\ \mathbf{d}_C,\ \mathbf{X} \rangle$, and let $t\left(\mathbf{Z}, \mathbf{e}, \mathbf{X}\right)$ be any function of $\mathbf{Z}$, $\mathbf{e}$ and $\mathbf{X}$. Because $\mathbf{e}$ and $\mathbf{X}$ are functions of $\mathcal{F}$, and randomization ensures $\Pr\left(\mathbf{Z} = \mathbf{z} \mid \mathcal{F}\right) = 1/\left|\Omega\right| = 1/L$, it follows that, for all v,

$$(1.1) \qquad \Pr\left\{t\left(\mathbf{Z}, \mathbf{e}, \mathbf{X}\right) \geq v \mid \mathcal{F}\right\} = \frac{\left|\{\mathbf{z} \in \Omega : t\left(\mathbf{z}, \mathbf{e}, \mathbf{X}\right) \geq v\}\right|}{\left|\Omega\right|},$$

which is the randomization distribution of $t\left(\mathbf{Z}, \mathbf{e}, \mathbf{X}\right)$. In words, given $\mathcal{F}$, the chance that $t\left(\mathbf{Z}, \mathbf{e}, \mathbf{X}\right) \geq v$ is simply the proportion of treatment assignments $\mathbf{z} \in \Omega$ such that $t\left(\mathbf{z}, \mathbf{e}, \mathbf{X}\right) \geq v$. Moreover, (1.1) is the distribution of $t\left(\mathbf{Z}, \mathbf{e}, \mathbf{X}\right)$ given $\mathcal{F}$ no matter what process produced $\mathcal{F}$. Fisher's (1935) description of randomization as the 'basis for inference' refers to the fact that randomization creates the distribution (1.1) for every function $\mathbf{e}$ of $\mathcal{F}$ without further assumptions.

The sharp *null hypothesis of no effect* asserts that the response of each individual is unchanged by receiving the treatment, $H_0 : \mathbf{r}_T = \mathbf{r}_C$. If H_0 were true, randomization labels people treated or control, but their depression is unchanged. If H_0 were true, the observed response $R_{ski} = Z_{ski}\,r_{Tski} + \left(1 - Z_{ski}\right) r_{Cski}$ equals r_{Cski}, or $\mathbf{R} = \mathbf{r}_C$.

If the null hypothesis of no effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, were true, then $\mathbf{R} = \mathbf{r}_C$, where $\mathbf{r}_C$ is a function of $\mathcal{F}$, so the null randomization distribution of $t\left(\mathbf{Z}, \mathbf{R}, \mathbf{X}\right) = t\left(\mathbf{Z}, \mathbf{r}_C, \mathbf{X}\right)$ is given by (1.1) with $\mathbf{e} = \mathbf{r}_C = \mathbf{R}$, where both $t\left(\mathbf{Z}, \mathbf{r}_C, \mathbf{X}\right)$ and its null distribution (1.1) can be calculated from the observed data when H_0 is true. For instance, in completely randomized experiments, Welch (1937) tested $H_0 : \mathbf{r}_T = \mathbf{r}_C$ using the randomization distribution of a test statistic suggested by analysis of variance, Gail,

Tan, and Piantadosi (1988) used the randomization distribution of a test statistic suggested by a generalized linear model, and Raz (1990) used the randomization distribution of a test statistic that adjusted for $\mathbf{X}$ using a data smoother, and Rosenbaum (2002) reviewed a class of randomization based covariance adjustments. In group randomized trials, Gail, et al. (1996) and Braun and Feng (2001) test no effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, using the group-randomization distribution (1.1) of a test statistic suggested by a generalized linear model. In some of these cases, the form of the test statistic is suggested by a model for parts of $\mathcal{F}$, but for a test of no effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, the significance level from (1.1) has the correct level whether or not that model actually generated $\mathcal{F}$.

Having tested the null hypothesis of no effect, it is natural to draw inferences about the magnitude of effect. Gail, et al. (1996) and Braun and Feng (2001) do this using a generalized linear model for the process that generates $\mathcal{F}$, so the parameters that define the magnitude of effect are now parameters of the model, and the associated confidence statements have the correct levels if the model is true. This is an entirely reasonable approach, but it is done "at the expense of modelling assumptions," to use Gail, et al.'s (1996, p. 1083) description. If the model is false, then the parameters that define the nonnull effect are not well defined. In particular, a stochastic model for the process that generates $\mathcal{F}$, say a logit model in which the treatment effect is identified with a log-odds ratio as in Gail, et. al. (1996, p. 1083), would not generally determine $\mathbf{r}_C$ from $\mathbf{R}$, and so one could generally calculate neither $t(\mathbf{Z}, \mathbf{r}_C, \mathbf{X})$ nor (1.1) with $\mathbf{e} = \mathbf{r}_C$. This says that these inferences about hypotheses other than no effect are valid inferences assuming the model for $\mathcal{F}$ is true, but they are no longer strict randomization inferences based on (1.1).

The approach we take is different in the following sense. Like Gail, et al. (1996) and Braun and Feng (2001), we test no treatment effect, $H_0 : \mathbf{r}_T = \mathbf{r}_C$, using (1.1). Unlike these authors, and paralleling more closely the literature on randomization inference in randomized experiments (e.g., Lehmann 1986, §5.14, pp. 245-248), we consider null hypotheses about treatment effects that relate $\mathbf{r}_T$ and $\mathbf{r}_C$ in such a way that if the null hypothesis were true, then the unobservable $\mathbf{r}_C$ could be calculated from the observed $\mathbf{R}$ *and* the null hypothesis. Having calculated $\mathbf{r}_C$ under the null hypothesis, we test exactly at level α the null hypothesis using (1.1). This approach will falsely reject a true null hypothesis with probability at most α with no assumption about a stochastic model that generated $\mathcal{F}$. To emphasize, there is no stochastic model for r_{Cski}, and no appeal to asymptotics. Specifically, §2.1 discusses hypotheses of constant effects as in Lehmann (1986, §5.14), §2.2 discusses effects in the presence of noncompliance, and §4 discusses dilated effects.

2. Randomization Inference When the Treatment Has an Effect

2.1. **Hypotheses of Constant Effect.** The *hypothesis of a constant treatment effect* says

$$(2.1) \qquad r_{Tski} - r_{Cski} = \tau, \text{ for } i = 1, \ldots, n_{sk}, \ k = 1, \ldots, K_s, \ s = 1, \ldots, S,$$

or a shift of τ, that is, $\mathbf{r}_T = \mathbf{r}_C + \tau\,\mathbf{1}$, or $\mathbf{R} = \mathbf{r}_C + \tau\,\mathbf{Z}$, where $\mathbf{1}$ is an N-dimensional vector of 1's. To test $H_0 : \tau = \tau_0$, compute $A_{\tau_0,ski} = R_{ski} - \tau_0\,Z_{ski}$ or $\mathbf{A}_{\tau_0} = \mathbf{R} - \tau_0\,\mathbf{Z}$. If $H_0 : \tau = \tau_0$ were true, then $\mathbf{A}_{\tau_0} = \mathbf{r}_C = \mathbf{e}$, say, is a function of $\mathcal{F}$, so $t\,(\mathbf{Z}, \mathbf{A}_{\tau_0}, \mathbf{X}) = t\,(\mathbf{Z}, \mathbf{r}_C, \mathbf{X}) = t\,(\mathbf{Z}, \mathbf{e}, \mathbf{X})$ is compared to (1.1) computed with $\mathbf{e} = \mathbf{r}_C = \mathbf{A}_{\tau_0}$. The test is inverted to obtain a confidence interval for τ, and

a Hodges-Lehmann point estimate is obtained by equating $t(\mathbf{Z}, \mathbf{A}_{\tau_0}, \mathbf{X})$ to its null expectation and solving as closely as possible for the estimator $\widehat{\tau}$. In a completely randomized experiment using Wilcoxon's rank sum test, this yields the standard, exact randomization-based confidence interval for a shift and the Hodges-Lehmann point estimate; see Lehmann (1998, §2).

2.2. **Noncompliance Hypotheses: Effect Proportional to Dose Received.**
Treatment is assigned at random, but patients decide for themselves whether, and to what extent, to comply with treatment. The degree of compliance is subject to biases of self-selection, similar to those that arise in nonrandomized observational studies. For a clear and memorable example of the problems that can arise, see May, et al. (1981). One good solution to these problems uses the random *assignment* of treatments as an instrument for the treatment *actually* received; see Angrist, Imbens and Rubin (1996). The random assignment of treatments is untainted by the patient's willingness to comply with the assigned treatment, but the treatment the patient actually receives is determined by a combination of random assignment and compliance, so it is tainted by self-selection. The basic idea behind the method of instrumental variables is to use an instrument, here random assignment of treatments, to extract variation in the treatment actually received that is independent of unmeasured confounding variables that determine compliance. This basic idea is formalized below. The approach we take here uses randomization inference with an instrumental variable (Rosenbaum 1996, 1999a, §5; Greevy, et al. 2004, Imbens and Rosenbaum 2005); however, we use the group randomization (1.1) together with patient level compliance.

An alternative to (2.1) says *effect on response is proportional to effect on dose of treatment*, that is,

$$(2.2) \qquad r_{Tski} - r_{Cski} = \beta \left(d_{Tski} - d_{Cski} \right), \text{ for all } ski;$$

see Rosenbaum (1996; 1999a, §5). This hypothesis says the treatment works solely by manipulating the dose received, a statement that is sometimes called the *exclusion restriction*. In the Prospect Study, $d_{Tski} - d_{Cski}$ is either $1 - 0 = 1$ for a complier or $0 - 0 = 0$ for a noncomplier, so the model says $r_{Tski} - r_{Cski} = \beta$ for a complier and $r_{Tski} - r_{Cski} = 0$ for a noncomplier; that is, a patient who does not accept treatment from the depression care manager receives no benefit from being at a practice with a depression care manager. If a depression care manager benefited depressed patients who did not accept care from the manager, then (2.2) would be false. This could happen if the manager affects the behavior of primary care physicians at the practice.

To test $H_0 : \beta = \beta_0$ in (2.2), compute $A_{\beta_0, ski} = R_{ski} - \beta_0 \, D_{ski}$ or $\mathbf{A}_{\beta_0} = \mathbf{R} - \beta_0 \, \mathbf{D}$. In the Prospect Study, $d_{Cski} = 0$, so if $H_0 : \beta = \beta_0$ is true in (2.2), then $\mathbf{A}_{\beta_0} = \mathbf{r}_C = \mathbf{e}$, say, so $t \left(\mathbf{Z}, \mathbf{A}_{\beta_0}, \mathbf{X} \right) = t \left(\mathbf{Z}, \mathbf{r}_C, \mathbf{X} \right) = t \left(\mathbf{Z}, \mathbf{e}, \mathbf{X} \right)$, and (1.1) is computed with $\mathbf{e} = \mathbf{r}_C = \mathbf{A}_{\beta_0}$.

This argument works in general, without the special structure that $d_{Cski} = 0$ in the Prospect Study. Consider the outcome (a_{Tski}, a_{Cski}) given by $a_{Tski} = r_{Tski} - \beta \, d_{Tski}$ and $a_{Cski} = r_{Cski} - \beta \, d_{Cski}$, so that (2.2) implies this outcome is unaffected by treatment, $a_{Tski} = a_{Cski} = a_{ski}$, say. Write $\mathbf{a}$ the vector of a_{ski}'s and notice that $\mathbf{a}$ is a function of $\mathcal{F}$. If $H_0 : \beta = \beta_0$ is true in (2.2), then $\mathbf{A}_{\beta_0} = \mathbf{R} - \beta_0 \, \mathbf{D} = \mathbf{a} = \mathbf{e}$, say, so $t \left(\mathbf{Z}, \mathbf{A}_{\beta_0}, \mathbf{X} \right) = t \left(\mathbf{Z}, \mathbf{a}, \mathbf{X} \right) = t \left(\mathbf{Z}, \mathbf{e}, \mathbf{X} \right)$ is compared with a randomization

distribution (1.1) with $\mathbf{e} = \mathbf{a} = \mathbf{A}_{\beta_0}$. See Rosenbaum (1999a, §5) for an example in which d_{Tski} and d_{Cski} are continuous.

2.3. **Covariance Adjustment in Randomization Tests.** The discussion in §2.1 and §2.2 made no use of the covariates, $\mathbf{X}$, but it is straightforward to incorporate them, with no change in the logic; see Rosenbaum (2002). Consider testing the hypothesis $H_0 : \tau = \tau_0$ in (2.1) with covariance adjustment. If $H_0 : \tau = \tau_0$ were true, then $\mathbf{A}_{\tau_0} = \mathbf{R} - \tau_0\,\mathbf{Z} = \mathbf{r}_C$ would be a function of $\mathcal{F}$, and $\mathbf{X}$ is a function of $\mathcal{F}$, so any function of $(\mathbf{A}_{\tau_0}, \mathbf{X})$ would also be a function of $\mathcal{F}$. One specific function of $(\mathbf{A}_{\tau_0}, \mathbf{X})$ is the vector, say $\mathbf{e}$, of residuals when $\mathbf{A}_{\tau_0}$ is regressed on $\mathbf{X}$ by any method of regression. Here, the regression is a fit, not a model; that is, it is simply a function of data $(\mathbf{A}_{\tau_0}, \mathbf{X})$ that produces residuals, $\mathbf{e}$, with no assumption that the regression is in any way related to whatever process produced $\mathcal{F}$. If $H_0 : \tau = \tau_0$ were true, then these residuals, $\mathbf{e}$, are functions of $(\mathbf{A}_{\tau_0}, \mathbf{X})$, where $\mathbf{A}_{\tau_0} = \mathbf{R} - \tau_0\,\mathbf{Z} = \mathbf{r}_C$ and $\mathbf{X}$ are functions of $\mathcal{F}$, so the randomization distribution of $t\,(\mathbf{Z}, \mathbf{e}, \mathbf{X})$ would be given by (1.1). Inverting this test yields a confidence set for an additive treatment effect τ. A parallel argument works for β in (2.2).

3. Exact Inference in Group Randomized Trials

3.1. **A Test Statistic and its Null Distribution in Group Randomized Trials.** Define $u_{ski\ell j} = 1$ if $e_{ski} > e_{s\ell j}$, $u_{ski\ell j} = -1$ if $e_{ski} < e_{s\ell j}$, and $u_{ski\ell j} = 0$ if $e_{ski} = e_{s\ell j}$, and calculate

$$W = \sum_{s=1}^{S} \frac{1}{n_s + 1} \sum_{k=1}^{K_s} \sum_{i=1}^{n_{sk}} \sum_{\ell=1}^{K_s} \sum_{j=1}^{n_{s\ell}} Z_{ski}\,(1 - Z_{s\ell j})\,u_{ski\ell j} = \sum_{s=1}^{S} \sum_{k=1}^{K_s} Z_{sk1}\,q_{sk+}$$

where $q_{ski} = (n_s + 1)^{-1} \sum_{\ell=1}^{K_s} \sum_{j=1}^{n_{s\ell}} u_{ski\ell j}$ and $q_{sk+} = \sum_{i=1}^{n_{sk}} q_{ski}$, using $Z_{ski} = Z_{sk1}$ for $i = 1, ..., n_{sk}$ and the fact, due to Mantel (1967), that $0 = \sum_{k=1}^{K_s} \sum_{i=1}^{n_{sk}} \sum_{\ell=1}^{K_s} \sum_{j=1}^{n_{s\ell}} Z_{ski} Z_{s\ell j} u_{ski\ell j}$, because $Z_{ski} Z_{s\ell j} u_{ski\ell j} = -Z_{s\ell j} Z_{ski} u_{s\ell jki}$ both appear in this sum and they cancel. Here, W is essentially a weighted sum of S Mann-Whitney-Wilcoxon statistics using van Elteren's (1960) optimal weights $1/(n_s + 1)$; see Noether (1963), Lehmann (1998, §3.3) and Rosner, et al. (2003, 2006). The validity of the randomization test is not affected by the choice of weights, but its power is affected, and the weights $1/(n_s + 1)$ are nearly optimal when the treatment effect is small and the intracluster correlation is low. More precisely, as can be quickly seen from the discussion in Noether (1963, §1), these standard weights would be proportional to weighting stratum specific results inversely as their variances if: (i) the null hypothesis of no effect were true, and (ii) if, contrary to fact, within strata, the clusters themselves had been formed by random assignment. See §5 for more about weights. When the null hypothesis is true, the group-randomization distribution (1.1) of W has $E(W) = 0$ and $var(W) = \sum_{s=1}^{S} \{m_s (K_s - m_s)\} / \{K_s (K_s - 1)\} \sum_{k=1}^{K_s} q_{sk+}^2$. The exact null distribution (1.1) is easily determined by direct calculation.

3.2. **Constant Effects in the Prospect Study.** The last two columns of Table 1 give the ranks q_{sk+}, $s = 1, \ldots, 10$, $k = 1, 2$, for testing one specific hypothesis, namely the hypothesis of no effect on the change in Hamilton Depression Scores, $H_0 : r_{Tski} = r_{Cski}$, after adjustment for Age and Suicidal Ideation. These were obtained by regressing the decline R_{ski} in Hamilton Scores on Age and Suicidal Ideation using M-estimation, obtaining the $N = 487$ residuals, **e**. The Splus defaults were used; see Rosenbaum (2007, §3.2) for discussion of the choice of ψ-function in m-testing. The statistic W is the sum of the scores in the second column, $W = -22.44$.

As in §2.3 and §3.1, if $H_0 : r_{Tski} = r_{Cski}$ were true, then the residual vector $\mathbf{e}$ would be a function of $\mathcal{F}$, so the randomization distribution of W would be given by (1.1), which simply permutes the ranks in Table 1 in all $2^{10} = 1024$ ways corresponding with the 2^{10} assignments $\mathbf{z} \in \Omega$, and of these, 8 produce values of W less than or equal to -22.44, so the one sided significance level is $8/1024 = 0.0078$.

The hypothesis $H_0 : \tau = -\frac{1}{2}$ in (2.1) is tested by the procedure applied to $\mathbf{A}_{\tau_0} = \mathbf{R} + \left(\frac{1}{2}\right) \mathbf{Z}$, yielding a new regression, residuals and scores q_{sk+}. The statistic is then $W = -19.40$ and 15 of the 1024 rearrangements of the new scores yield smaller values of W, so the exact one-sided significance level is $15/1024 = 0.0146$. For $H_0 : \tau = -1.508$ the one-sided significance level is $51/1024 = 0.0498$, whereas for $H_0 : \tau = -1.509$ the one-sided significance level is $52/1024 = 0.0508$, and the one-sided, exact 95% confidence interval is $\tau \leq -1.509$. The Hodges-Lehmann point estimate $\widehat{\tau}$ of τ in (2.1) equates W to its null expectation, namely 0, and solves for $\widehat{\tau}$. For $H_0 : \tau = -3.258$ the statistic is $W = -0.068$, whereas with $H_0 : \tau = -3.259$ the statistic is $W = 0.085$, so $\widehat{\tau} \doteq -3.26$.

Figure 2 displays the residuals e_{ski} from the constant effect model (2.1), obtained by regressing $\mathbf{A}_{\widehat{\tau}} = \mathbf{R} - \widehat{\tau} \mathbf{Z}$ with $\widehat{\tau} = -3.26$ on $\mathbf{X}$ using M-estimation. Here, $\mathbf{e}$ is plotted against the treatment group, $\mathbf{Z}$, for 487 patients. The two boxplots look similar, which is consistent with an additive effect of $\widehat{\tau} = -3.26$.

3.3. **Alternative Methods of Covariance Adjustment.** The current section uses an additional covariate, namely the baseline Hamilton Score, and compares several methods of covariance adjustment, including M-estimation, least squares, generalized additive models (Buja, Hastie, Tibshirani 1989) and projection pursuit (Friedman and Stuetzle 1981). Unlike the other methods, for projection pursuit,

the inference about τ in (2.1) is different depending upon whether the outcome is level of the Hamilton Score or the change from baseline, and both are considered. The default implementations in Splus are used, with loess used as the smoother for baseline score and Age in the generalized additive models.

Table 2 compares five methods of covariance adjustment in testing three hypotheses about a constant effect, $H_0 : \tau = \tau_0$, in (2.1). Exact, one-sided significance levels are given using (1.1). The results are similar with the exception of projection pursuit performed on changes, which produced larger significance levels.

3.4. **Imperfect Compliance.** Unlike (2.1), in (2.2) a patient benefits only if the patient makes use of the depression care manager. To test $H_0 : \beta = \beta_0$ in (2.2), one computes $\mathbf{A}_{\beta_0} = \mathbf{R} - \beta_0 \mathbf{D}$, which is a function of $\mathcal{F}$ if H_0 is true, and then proceeds exactly as in §3.2. The one-sided significance level for testing $H_0 : \beta = 0$ is the same as for testing $H_0 : \tau = 0$ or $H_0 : \mathbf{r}_T = \mathbf{r}_C$, namely $8/1024 = 0.0078$, because it is the exactly same hypothesis, tested by the same test W, compared to the same null randomization distribution (1.1).

If (2.2) were true, the effect on compliers would have a one-sided exact 95% confidence interval of $\beta \leq -1.70$ and a Hodges-Lehmann point estimate of $\widehat{\beta} = -3.49$. Although the estimated effect for compliers is marginally larger than the constant effect, the difference is quite small, in part because compliance was 93% overall; see Table 1. Residuals e_{ski} for the effect-proportional-to-dose of treatment model (2.2) were obtained by regressing $\mathbf{A}_{\widehat{\beta}} = \mathbf{R} - \widehat{\beta} \mathbf{D}$ on $\mathbf{X}$ using M-estimation with $\widehat{\beta} = -3.49$. Boxplots (not shown) of these residuals e_{ski} resemble Figure 2 and are compatible with (2.2) with $\widehat{\beta} = -3.49$.

4. Different Effects at Different Quantiles

In place of (2.1), consider the following model for the treatment effect, (r_{Tski}, r_{Cski}),

$$(4.1) \qquad r_{Cski} = r_{Tski} + \Delta(r_{Tski}), \ \ with \ \Delta(r) \geq 0, \ \Delta(r) \geq \Delta(r') \ for \ r \geq r',$$

so $\Delta(\cdot)$ is nonnegative and nondecreasing; see Rosenbaum (1999b) where this is called a dilated effect. Here, (4.1) says that the depression care manager is never harmful, as $\Delta(r_{Tski}) \geq 0$ for all ski, but the benefits $\Delta(r_{Tski})$ are greater (or at least as great) when r_{Tski} is larger, so patients with higher values of (r_{Tski}, r_{Cski}) benefit most from the depression care manager. Although (4.1) describes the unobservable joint behavior of (r_{Tski}, r_{Cski}), it may be reinterpreted as a statement about the observable marginal distributions (R_{ski}, Z_{ski}); see Doksum and Sievers (1976).

Some evidence for varied effects consistent with (4.1) appears in Figure 3, which is a quantile-quantile plot of changes in Hamilton Scores in the control and treated groups. Two lines appear in Figure 3, a dotted line for $y = x$ and a solid line for $y = x + 3.26$, where $\widehat{\tau} = -3.26$ in §3.2. Ignoring the issue of group randomization for a moment, one would expect the points to fall roughly on the dotted line if the treatment had no effect, and roughly on the solid line if the treatment had a constant effect of $\widehat{\tau} = -3.26$. Figure 3 suggests that the effect may be larger at upper quantiles and smaller at lower quantiles. Table 3 displays several quantiles of changes in depression scores in treated and control groups together with the difference in those quantiles. In the context of the Prospect Study, a plot like Figure 3 might arise if the typical patient benefits by about $\widehat{\tau} = -3.26$ from the presence of a depression care manager, but patients who would have gotten substantially worse

under standard care benefited by more than $\widehat{\tau} = -3.26$, whereas patients who would have improved markedly under standard care benefited by less than $\widehat{\tau} = -3.26$.

Let $r_{C(1)} \leq \ldots \leq r_{C(N)}$ and $r_{T(1)} \leq \ldots \leq r_{T(N)}$ be the unobserved, ordered potential responses of all $N = 487$ subjects, which are functions of $\mathcal{F}$. Because only some of the r_{Cski}'s are observed, none of the $r_{C(j)}$ are observed, and similarly for the $r_{T(j)}$. Fix an integer f, $1 \leq f \leq N$, and write $\rho = r_{T(f)}$, and $\Delta_\rho = \Delta(\rho)$, so ρ and Δ_ρ are functions of $\mathcal{F}$. In words, $\rho = r_{T(f)}$ is the f/N quantile of changes in depression scores that would have been observed had all patients been in a practice with a depression care manager, and Δ_ρ is the effect at this quantile.

For a specific f, such as the $f/N \doteq \frac{3}{4}$ for the upper quartile, consider testing the hypothesis $H_0 : \Delta_\rho = \Delta_{0\rho}$ for some specified number $\Delta_{0\rho}$. Write $sign(y) = 1$ if $y > 0$, $sign(y) = 0$ if $y = 0$, $sign(y) = -1$ if $y < 0$. It is easy to verify (e.g. Rosenbaum 1999b) that (4.1) implies $sign\{R_{ski} - (1 - Z_{ski})\Delta_\rho - \rho\} = sign(r_{Tski} - \rho)$, where $sign(r_{Tski} - \rho)$ is a function of $\mathcal{F}$; here, $R_{ski} - (1 - Z_{ski})\Delta_\rho - \rho$ is *not* a function of $\mathcal{F}$, but its *sign* is a function of $\mathcal{F}$. Sort the adjusted responses, $A_{ski}^{\Delta_{0\rho}} = R_{ski} - (1 - Z_{ski})\Delta_{0\rho}$, yielding order statistics, $A_{(1)}^{\Delta_{0\rho}} \leq \ldots \leq A_{(N)}^{\Delta_{0\rho}}$, and write $e_{ski}^{\Delta_{0\rho}} = 1$ if $A_{ski}^{\Delta_{0\rho}} \geq A_{(f)}^{\Delta_{0\rho}}$, $e_{ski}^{\Delta_{0\rho}} = 0$ otherwise. If $H_0 : \Delta_\rho = \Delta_{0\rho}$ is true, then $sign\left\{A_{ski}^{\Delta_{0\rho}} - \rho\right\} = sign(r_{Tski} - \rho)$ and $A_{(f)}^{\Delta_{0\rho}} = \rho$, and $e_{ski}^{\Delta_{0\rho}} = 1$ if $sign(r_{Tski} - \rho) \geq 0$, $e_{ski}^{\Delta_{0\rho}} = 0$ otherwise, so $e_{ski}^{\Delta_{0\rho}}$ is a function of $\mathcal{F}$. In cluster sk, the proportion of individuals with $A_{ski}^{\Delta_{0\rho}} \geq A_{(f)}^{\Delta_{0\rho}}$ is $q_{sk}^{\Delta_{0\rho}} = (1/n_{sk})\sum_{i=1}^{n_{sk}} e_{ski}^{\Delta_{0\rho}}$, which is a function of $\mathcal{F}$ if $H_0 : \Delta_\rho = \Delta_{0\rho}$ is true. With paired clusters, $K_2 = 2$, $m_s = 1$, as in the Prospect Study, the treated-minus-control difference in proportions in stratum s is $\sum_{k=1}^{2} Z_{sk1} q_{sk}^{\Delta_{0\rho}} - \sum_{k=1}^{2}(1 - Z_{sk1}) q_{sk}^{\Delta_{0\rho}} = \sum_{k=1}^{2} Z_{sk1} 2\left(q_{sk}^{\Delta_{0\rho}} - \overline{q}_s^{\Delta_{0\rho}}\right)$ where $\overline{q}_s^{\Delta_{0\rho}} = (1/2)\sum_{k=1}^{2} q_{sk}^{\Delta_{0\rho}}$. The Cochran-Mantel-Haenszel (CMH) test (e.g.,

Fleiss, Leven and Paik 2003, p. 253) attaches weights $n_{s1}\, n_{s2}/n_s$ to these differences in proportions, yielding the statistic $H = \sum_{s=1}^{S} \sum_{k=1}^{2} Z_{sk1} \left(\frac{2\, n_{s1}\, n_{s2}}{n_s} \right) \left(q_{sk}^{\Delta_{0\rho}} - \overline{q}_s^{\Delta_{0\rho}} \right)$ whose exact group-randomization distribution (1.1) is easily determined. In parallel with §3.1, the CMH weights, $n_{s1}\, n_{s2}/n_s$, approximate optimal weights for small effects when the intracluster correlation is low.

In the Prospect Study, there is just a little evidence of a dilated effect. The one-sided 95% confidence interval for τ in §3.2 suggested an effect of at least 1.509 points. The 95% interval for Δ_ρ includes 1.509 points for $f = 48, 122, 244,$ and 366, that is, for the lower 10% quantile, the lower quartile, the median and the upper quartile. Only at $f = 438$ for the upper 10% quantile is the one sided 95% interval shorter for Δ_ρ; specifically it is $\Delta_\rho \geq 3.33$, with Hodges-Lehmman point estimate $\widehat{\Delta}_\rho = 6.0$, found by setting $H = 0$ and solving for $\widehat{\Delta}_\rho$. This suggests that for the 10% of patients who would experience the smallest improvements, the effect of the depression care manager might possibly be twice as large as for the typical patient.

5. A Simulation of Power

The weighted Wilcoxon statistic, W, in §3.1 used van Elteren's (1960) weights which would be optimal if the intra-cluster correlation were zero. Because W is a randomization test, the test has the correct level from (1.1) whether or not these are the ideal weights, but the choice of weights does affect the power of the test. Do van Elteren's (1960) weights yield reasonable power when the clustering does matter?

The size and power of the randomization test based on W were compared to size and power of the Wald test from fitting a linear mixed model. The linear mixed model (LMM) had an additive fixed treatment effect, fixed pair effects, random

Normal cluster effects, and random Normal patient errors; it was sometimes the correct model, sometimes an incorrect model because the true distributions were not Normal. With ten paired clusters, as in the Prospect Study, the Wald test statistic was compared to the t-distribution with 9 degrees of freedom, as suggested by Feng, et al. (2001, p. 175). Of course, unlike W, the LMM test is not a randomization test, so its level is not guaranteed by random assignment of treatments — it may have the wrong level if the model is wrong.

In most simulated situations, as in the Prospect Trial itself, there were ten pairs, $s = 1, \ldots, 10$ of two clusters, $k = 1, 2$, with one cluster picked at random for treatment within each pair. In some situations, the cluster sizes, n_{sk}, were the unbalanced cluster sizes in Table 1 from the Prospect Trial (labeled ACTUAL), in others they were DOUBLE the sizes in the trial, $2n_{sk}$, and in still others they were HALF the sizes rounded up to the nearest integer, $\lceil n_{sk}/2 \rceil$. The simulation also considered EQUAL cluster sizes, $n_{sk} = 25$ for all s, k. Finally, EQUAL IN PAIR means cluster sizes that varied between pairs but were constant within pairs, specifically in pair s both clusters had size $\lceil (n_{s1} + n_{s2})/2 \rceil$ where the n_{sk} are given in Table 1, so in pair $s = 2$, both clusters had size $\lceil (31 + 6)/2 \rceil = 18$. Data were generated by a linear mixed model of the form $R_{ski} = \tau Z_{ski} + \phi_s + \gamma_{sk} + \zeta_{ski}$. Here, $Z_{ski} = 1$ if cluster s, k was assigned to treatment, and $Z_{ski} = 0$ if the cluster was assigned to control, so τ is the constant treatment effect. For $\tau = 0$, the proportion of rejections of $H_0 : \tau = 0$ estimates the size of a test that aspires to have level 0.05. For $\tau = -2$, roughly similar to the Prospect Trial estimate, the proportion of rejections estimates power against this alternative. Also ϕ_s is a fixed pair effect, γ_{sk} is a random cluster effect,

and ζ_{ski} is a random patient effect, and all of the γ_{sk}'s and ζ_{ski}'s are mutually independent and independent of the random treatment assignments Z_{ski}. As indicated in Table 4, errors ζ_{ski} had one of three distributions: a Normal distribution with expectation zero and standard deviation 7, $N(0, 49)$, a scaled t-distribution with three degrees of freedom, or a Cauchy distribution. In looking at Table 4, keep in mind that a standard Cauchy distribution eventually has thicker tails than any Normal distribution, but is typically much less dispersed than a $N(0, 49)$. As indicated in Table 4, the clusters γ_{sk} were sometimes Normally distributed, sometimes standard Cauchy, sometimes just zero, labeled ZERO. We tried several other cases, such as clusters γ_{sk} that were exponentially distributed, but the results were qualitatively similar and are not reported. The pair effects ϕ_s were set to zero, because for both W and LMM, they do not affect the inference.

The intra-cluster correlation λ is $var(\gamma_{sk})/\{var(\gamma_{sk}) + var(\zeta_{ski})\}$ when γ_{sk} and ζ_{ski} both have finite variance, so the variance of the mean of n_{sk} responses from one cluster has variance $var(\gamma_{sk}) + var(\zeta_{ski})/n_{sk} = \{\lambda/(1-\lambda) + 1/n_{sk}\} var(\zeta_{ski})$. Hannan, et al (1994) report λ's ranging from 0.002 to 0.012 for various outcomes in the Minnesota Heart Health Community Trial. Feng, et al. (2001) report λ's between 0.01 and 0.03 for the Working Well Trial. In the Prospect Trial, using a linear mixed model, we estimated a λ of 0.028. In Table 4, by adjusting the scale of the cluster γ_{sk} distribution, λ's of 0, 0.02, 0.04, 0.08 and 0.25 were obtained, where $0.08 = 2.67 \times 0.03$ is more than two and a half times larger than the largest λ found for several outcomes in the several group randomized trials just mentioned. To put this in perspective, if $\lambda = 0.08$ and $n_{sk} = 12$, then $\lambda/(1-\lambda) > 1/n_{sk}$, so that more than half of the variance of the mean of $n_{sk} = 12$ patient responses in the same cluster

is from cluster effects. Arguably, if $\lambda = 0.25$, one should not do a group-randomized experiment, because additional observations from the same cluster are worth much less than additional observations from a new cluster; this shows up clearly in the simulation. With Cauchy distributions, λ does not exist.

Each situation was simulated 1000 times, using antithetic variates to boost simulation efficiency (i.e., reversing the cluster assigned to treatment in each pair). For $\tau = 0$, the estimated standard error of the estimated proportion of rejections is always less than 0.007, whereas for $\tau = -2$, it is always less than 0.013.

In Table 4, the level is close to the nominal 0.05 level for both procedures. The powers of W and LMM are remarkably similar, except in the case of Cauchy errors. Generally, LMM has a small edge with Normal errors, and W has a small edge with errors from a t-distribution with 3 degrees of freedom. With Cauchy errors, W is much better. Although the intra-cluster correlation λ strongly affects power, it affects W and LMM in a similar way. With a high intra-cluster correlation of $\lambda = 0.25$, doubling or halving the sample size within clusters has very slight effects on power, suggesting that money spent on obtaining data on additional patients within the same clusters is money spent unwisely.

We also tried doubling the number of clusters, with the same paired cluster sizes as in the Prospect Study, each pair appearing twice. With $|\Omega| = 2^{20}$, we used the large sample approximation to (1.1) based on the moments of W in §3.1. With $\lambda = 0.08$ and Normal distributions, the power was 0.677 for W and 0.715 for LMM.

In short, W and LMM performed similarly, with the only decisive advantage being the superior performance of W for Cauchy errors. The use of van Elteren's weights, which have optimal properties for $\lambda = 0$, did not seem to create major problems for

the relative performance of W and LMM for $0 \leq \lambda \leq 0.25$. The reported values of λ from clinical trials that we could find were all in $0 \leq \lambda \leq 0.03$.

6. SUMMARY

In a group randomized trial of treatments for depression, we used the group-randomization (1.1) as the basis for inference about the magnitude of effects on individuals in group randomized trials, without assuming a distributional model for the generation of $\mathcal{F}$. Hypothesized effects included constant effects, effects proportional to dose received in the case of noncompliance, and dilated effects with larger effects at upper quantiles. Covariance adjustments were performed using various methods with no need to postulate a 'true' covariance model.

We found evidence that the depression care manager in the Prospect Study was beneficial as did Bruce et al. (2004). Using covariance adjustment with M-estimation in the intent-to-treat analysis, the one-sided 95% confidence bound suggests benefits of at least 1.509 points on the change in Hamilton Depression Score from baseline to 4 months. Accounting for noncompliance, the effect on a compliant patient appears to be slightly larger, at least 1.70 points. There is a little evidence that the effect of the depression care manager is not constant, but dilated, with the depression care manager being more beneficial to those patients who would have gotten considerably worse under the control.

REFERENCES

[1] Angrist, J., Imbens, G., and Rubin, D. (1996), "Identification of causal effects using instrumental variables," *JASA*, **91**, 444-469.

[2] Braun, T. M. and Feng, Z. (2001), "Optimal permutation tests for the analysis of group randomized trials," *Journal of the American Statistical Association*, **96**, 1424-1432.

[3] Brookmeyer, R. and Chen, Y. Q. (1998), "Person-time analysis of paired community intervention trials when the number of communities is small," *Statistics in Medicine*, **17**, 2121-2132.

[4] 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.

[5] Buja, A., Hastie, T., Tibshirani, R. (1989), "Linear smoothers and additive models," *Annals of Statistics*, **17**, 453-510.

[6] Cornfield, J. (1978), "Randomization by group," *American Journal of Epidemiology*, **108**, 100-102.

[7] Cox, D. R. and Reid, N. (2000) *The Theory of the Design of Experiments*, New York: CRC.

[8] Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley.

[9] Doksum, K. and Sievers, G. (1976), "Plotting with confidence," *Biometrika*, **63**, 421-434.

[10] Donner, A. (1998), "Some aspects of the design and analysis of cluster randomized trials," *Applied Statistics*, **47**, 95-113.

[11] Feng, Z., Diehr, P., Peterson, D. and McLerran, D. (2001), "Selected statistical issues in group randomized trials," *Annual Review of Public Health*, **22**, 167-187.

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

[13] Fleiss, J. L., Levin, B., and Paik, M. C. (2003), *Statistical Methods for Rates and Proportions* (3rd edition), New York: Wiley.

[14] Frangakis, C. E., Rubin, D. B., and Zhou, X. H. (2002), "Clustered encouragement designs with individual level noncompliance," *Biostatistics*, **3**, 147-177/

[15] Friedman, J. H. and Stuetzle, W. (1981), "Projection pursuit regression," *JASA*, **76**, 817-823.

[16] Gail, M. H., Tan, W. Y., and Piantadosi, S. (1988), "Tests for no treatment effect in randomized clinical trials," *Biometrika*, **75**, 57-64.

[17] Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B., and Pee, D. (1996), "On design consid-erations and randomization-based inference for community intervention trials," *Statistics in Medicine*, **15**, 1069-1092.

[18] Greevy, R., Silber, J. H., Cnaan, A., and Rosenbaum, P.R. (2004), "Randomization inference with imperfect compliance in the AAA randomized trial," *JASA*, **99**, 7-15.

[19] Hannan, P., Murray, D., Jacobs, D. Jr., and McGovern, P. (1994), "Parameters to aid in the design and analysis of community trials," *Epidemiology*, 88-95.

[20] Imbens, G. and Rosenbaum, P. R. (2005), "Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education," *Journal of the Royal Statistical Society*, A, **168**, 109-126.

[21] Lehmann, E. L. (1986), *Testing Statistical Hypotheses* (2nd edition), New York: Wiley.

[22] Lehmann, E. L. (1998), *Nonparametrics*, New Jersey: Prentice Hall.

[23] Mantel, N. (1967), "Ranking arbitrarily restricted observations," *Biometrics*, **23**, 65–78.

[24] May, G. S., Chir, B., DeMets, D. L., Friedman, L. M., Furberg, C., and Passamani, E. (1981), "The randomized clinical trial: Bias in analysis," *Circulation*, **64**, 669-673.

[25] Murray, D. (1998), *Design and Analysis of Group Randomized Trials*, Oxford: New York.

[26] Murray, D., Hannan, P. J., Pals, S. P., McCowen, R. G.,Baker, W. L., and. Blitstein J. L. (2006), "A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial," *Statistics in Medicine*, **25**, 375-388.

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

[28] Noether, G. E. (1963), "Efficiency of the Wilcoxon two-sample statistic for randomized blocks," *JASA* 894-898.

[29] Pitman, E. J. G. (1937), "Significance tests which may be applied to samples from any popu-lation," *Journal of the Royal Statistical Society*, B, 4, 119-130.

[30] Rosenbaum, P. R. (1996), "Comment," *JASA* **91**, 465-468.

[31] Rosenbaum, P. R. (1999a) Using combined quantile averages in matched observational studies," *Applied Statistics* **48** 63-78.

[32] Rosenbaum, P. R. (1999b), "Reduced sensitivity to hidden bias at upper quantiles in observational studies with dilated treatment effects," *Biometrics*, **55**, 560-564.

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

[34] Rosenbaum, P. R. (2007), "Sensitivity analysis for M-estimates, tests and confidence intervals in matched observational studies," *Biometrics*, to appear.

[35] Rosner, B., Glynn, R. J., and Lee, M. T. (2003), "Incorporation of clustering effects for the Wilcoxon rank sum test," *Biometrics*, **59**, 1089-1098.

[36] Rosner, B., Glynn, R. J., and Lee, M. T. (2006), "Extension of the rank sum test for clustered data," *Biometrics*, **62**, 1251-1259.

[37] Raz, J. (1990), "Testing for no effect when estimating a smooth function by nonparametric regression: A randomization approach," *JASA,* **85**, 132-138.

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

[39] Rubin, D. B. (1986), "Which ifs have causal answers?" *JASA*, **81**, 961–962.

[40] van Elteren, P. H. (1960), "On the combination of independent two sample tests of Wilcoxon," *Bulletin de l'Institut International de Statistique*, **37**, 351-361.

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

TABLE 1. Ten Pairs of Two Practices, One Treated (T), the Other Control (C). For each practice, the values are: the sample size (n), the mean change in Hamilton Depression Score, the percent compliance in the treated practice, and the rank scores, q_{sk+}, for testing no effect.

Pair	C n	T n	C Mean	T Mean	Compliance (%)	C q_{sk+}	T q_{sk+}
1	44	49	−4.7	−4.6	88	−0.79	0.79
2	31	6	−0.3	−7.3	100	3.00	−3.00
3	5	27	−2.0	−8.2	85	1.61	−1.61
4	22	1	−3.7	−3.0	100	−0.33	0.33
5	29	26	−2.8	−7.7	92	4.21	−4.21
6	5	37	−6.6	−5.9	97	−0.26	0.26
7	29	17	−5.0	−9.9	100	4.32	−4.32
8	22	40	−4.7	−8.7	95	4.49	−4.49
9	23	20	−4.9	−9.1	95	4.00	−4.00
10	24	30	−5.9	−9.1	93	2.18	−2.18
All	253	234	−3.9	−7.5	93	22.44	−22.44

TABLE 2. Exact, One-sided Significance Levels with Five Types of Covariance Adjustment, for Three Null Hypotheses.

Method of Covariance Adjustment	$H_0 : \tau = 0$	$H_0 : \tau = -0.5$	$H_0 : \tau = -0.7$
M-estimation	0.0127	0.0342	0.0479
Least Squares	0.0186	0.0352	0.0479
Generalized Additive	0.0176	0.0342	0.0537
Projection Pursuit; Changes	0.0303	0.0527	0.0605
Projection Pursuit; Levels	0.0146	0.0283	0.0430

TABLE 3. Observed Quantiles of Changes in Depression Scores By Treatment Group.

Quantile	$\frac{1}{10}$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{9}{10}$
Control	−13.9	−9.0	−4.0	0.9	7.0
Treated	−16.0	−12.0	−7.0	−3.0	1.0
Difference	2.1	3.0	3.0	3.9	6.0

TABLE 4. Randomization Test (W) versus Linear Mixed Model (LMM), for several intra-cluster correlations (λ).

| Treatment (Effect τ)	Clusters	Cluster Sizes	Errors	λ	LMM Power	W Power
0	N(0,1)	Actual	N(0,49)	0.02	0.049	0.057
0	N(0,1)	Equal in Pair	N(0,49)	0.02	0.034	0.037
0	Cauchy	Actual	N(0,49)	-	0.039	0.051
-2	Zero	Actual	N(0,49)	0	0.805	0.798
-2	N(0,1)	Actual	N(0,49)	0.02	0.688	0.679
-2	N(0,1)	Actual	$\frac{7}{\sqrt{3}}$t(3)	0.02	0.705	0.793
-2	N(0,1)	Equal	N(0,49)	0.02	0.739	0.738
-2	N(0,1)	Double	N(0,49)	0.02	0.835	0.798
-2	N(0,1)	Half	N(0,49)	0.02	0.492	0.500
-2	N(0,1)	Equal in Pair	N(0,49)	0.02	0.751	0.724
-2	N(0,2.04)	Actual	N(0,49)	0.04	0.574	0.556
-2	N(0,2.04)	Actual	$\frac{7}{\sqrt{3}}$t(3)	0.04	0.601	0.642
-2	N(0,2.04)	Equal	N(0,49)	0.04	0.654	0.648
-2	N(0,2.04)	Double	N(0,49)	0.04	0.685	0.637
-2	N(0,2.04)	Half	N(0,49)	0.04	0.426	0.425
-2	N(0,2.04)	Equal in Pair	N(0,49)	0.04	0.658	0.629
-2	N(0,3.13)	Actual	N(0,49)	0.06	0.509	0.459
-2	N(0,3.13)	Actual	$\frac{7}{\sqrt{3}}$t(3)	0.06	0.510	0.524
-2	N(0,3.13)	Double	N(0,49)	0.06	0.579	0.545
-2	N(0,3.13)	Half	N(0,49)	0.06	0.402	0.370
-2	N(0,3.13)	Equal in Pair	N(0,49)	0.06	0.556	0.534
-2	N(0,4.26)	Actual	N(0,49)	0.08	0.455	0.441
-2	N(0,4.26)	Actual	$\frac{7}{\sqrt{3}}$t(3)	0.08	0.440	0.448
-2	N(0,4.26)	Double	N(0,49)	0.08	0.515	0.460
-2	N(0,4.26)	Half	N(0,49)	0.08	0.365	0.347
-2	N(0,4.26)	Equal in Pair	N(0,49)	0.08	0.489	0.475
-2	N(0,1)	Actual	Cauchy	-	0.176	0.946
-2	N(0,1)	Double	Cauchy	-	0.152	0.949
-2	N(0,1)	Equal in Pair	Cauchy	-	0.220	0.968
-2	Cauchy	Actual	N(0,49)	-	0.226	0.283
-2	N(0,16.33)	Actual	N(0,49)	0.25	0.220	0.204
-2	N(0,16.33)	Actual	$\frac{7}{\sqrt{3}}$t(3)	0.25	0.231	0.213
-2	N(0,16.33)	Double	N(0,49)	0.25	0.257	0.207
-2	N(0,16.33)	Half	N(0,49)	0.25	0.224	0.202
-2	N(0,16.33)	Equal in Pair	N(0,49)	0.25	0.252	0.233
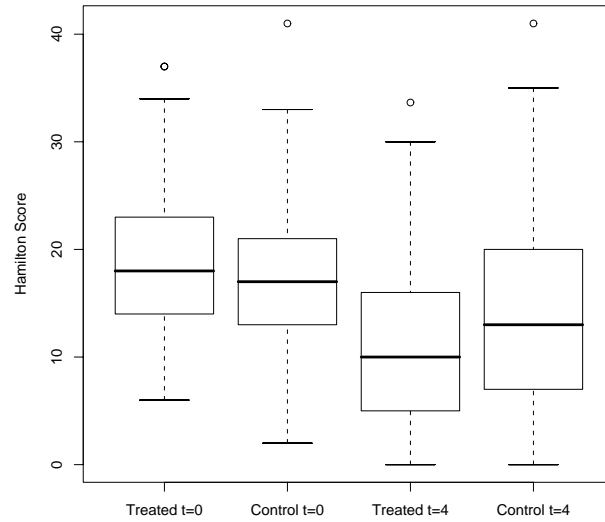
Figure 1: Hamilton Depression Scores by Assigned Treatment Group for 253 Patients in Treated Practices and 234 Patients in Control Pratices, at Baseline Before Treatment (*t=0*) and at Four Months (*t=4*).
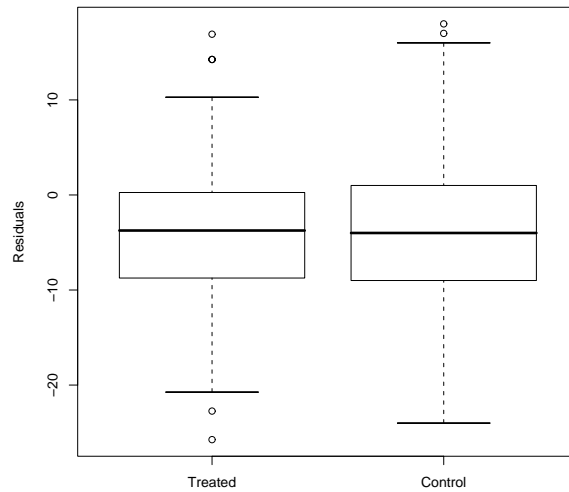

Figure 2: Residuals from Constant Effect Model: Boxplots of 487 Residuals e_{ski} from the m-Estimate Regression of $\mathbf{R} - \tau_0 \mathbf{Z}$ on $\mathbf{X}$, with $\tau_0 = -3.26$, for Treated $Z_{ski} = 1$ and Control $Z_{ski} = 0$ Groups.
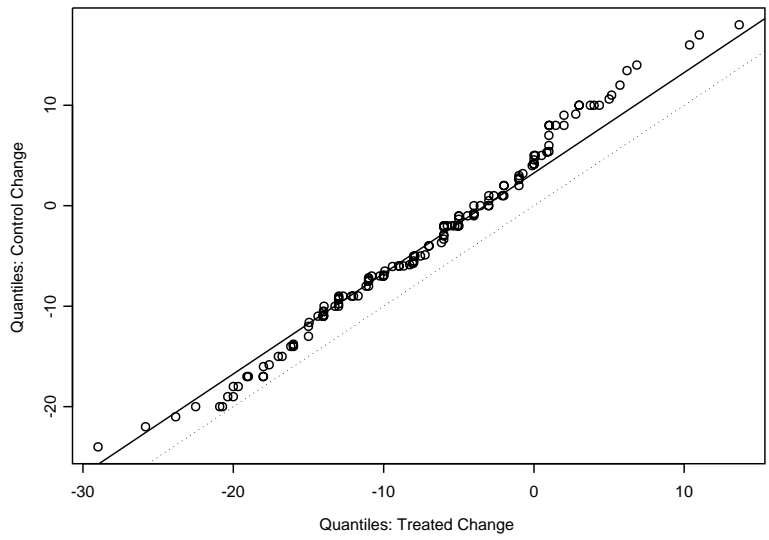
Figure 3: Quantile-Quantile Plot of Changes in Hamilton Scores R_{ski} in Treated ($Z_{ski} = 1$) and Control ($Z_{ski} = 0$) Groups. Dotted Line Is $y = x$. Solid Line Is $y = x + 3.26$.