Randomization Inference With Imperfect Compliance in the ACE-Inhibitor After Anthracycline Randomized Trial

Robert GREEVY, Jeffrey H. SILBER, Avital CNAAN, and Paul R. ROSENBAUM

Anthracyclines are quite effective at curing certain cancers of childhood, but they may damage the heart. The ACE-Inhibitor After Anthracycline (AAA) study compared enalapril to placebo in a randomized trial in an effort to determine whether treatment with enalapril would preserve or improve cardiac function among children previously treated with anthracylines. As is true in many clinical trials, patient compliance with the study protocol was imperfect; some children took less than the prescribed dose of enalapril or placebo. Most analytical procedures that acknowledge imperfect compliance do so at significant cost, abandoning the tight logic of random assignment. With noncompliance, assignment to enalapril or placebo is randomized, but the dose of enalapril actually received is not, and self-selection effects parallel to those in observational studies can exist and have been documented in some instances. Some researchers advocate adherence to the strict logic of randomization by reporting only, or else strongly emphasizing, the so-called "intent-to-treat" analysis, which makes no use of information about compliance. Other researchers report analyses that are not justified by random assignment and can be subject to substantial biases, such as "per protocol" analyses or "treatment received" analyses. Here we apply a recent proposal for randomization inference with an instrumental variable that uses randomization as the "reasoned basis for inference" in Fisher's phrase. We make no assumption that compliance is random; indeed, compliance may be severely biased. Importantly, the proposed analysis will find a statistically significant effect of the treatment if and only if the intent-to-treat analysis finds a significant effect; yet, unlike intent-to-treat analysis, our analysis acknowledges that a patient assigned to a drug that he or she does not take will not receive the drug's pharmacological benefits.

KEY WORDS: Clinical trial; Hodges–Lehmann estimate; Instrumental variable; Noncompliance; Permutation test; Randomization test; Randomized experiment; Randomized trial.

1. THE TRIAL, RANDOMIZATION, INSTRUMENTS, AND OUTLINE

1.1 The AAA Randomized Trial: Cardiac Damage After Chemotherapy

Although anthracyclines are quite effective at curing certain childhood cancers, with cure rates of perhaps two-thirds, it is estimated that more than half of the long-term survivors will show cardiac abnormalities by ecocardiology or angiography 10–20 years after diagnosis. The ACE-Inhibitor After Anthracycline (AAA) study (Silber et al. 2001, 2004) examined children under age 20 who had survived at least 4 years after cancer diagnosis and at least 2 years after the completion of all cancer treatment, and who had certain defined forms of decline in cardiac systolic performance after treatment with an anthracycline. It was hoped that the treatment, enalapril, would prevent further cardiac function reduction among such children. A total of 135 children were randomly assigned to enalapril or placebo, with cardiac performance measured at baseline, before treatment, and then at 6-month intervals thereafter. The trial began in October 1994, closed enrollment in March 1999, and collected data until March 2001; so the final patient to be enrolled could be followed for 2 years, and many patients were observed for much longer.

In many, if not most, clinical trials, some patients do not comply fully with the treatment to which they are assigned. For instance, a patient may consume only a fraction of the assigned dose of drug or even none at all. In the AAA study, the children were asked at regular intervals about pill consumption in the previous week, and some children consumed much less than was prescribed. Common sense suggests that a drug is unlikely to work if it remains in the bottle.

Common sense also suggests that patients who comply with the protocol may differ from those who refuse. Assignment to treatment or placebo is random, truly random, producing comparable groups, but compliance is a choice, not an accident. At the same time that children are choosing to accept or refuse medication, they are also choosing to exercise or watch TV, to join the football team or take cocaine, and there is nothing to ensure that compliers are comparable to noncompliers in ways that matter for cardiac performance. Perhaps this was demonstrated most vividly by May et al. (1981) in their discussion of the Coronary Drug Project, a randomized trial of lipid-lowering drugs for individuals who had survived myocardial infarction. Good compliers in the group assigned to clofibrate had a 5-year mortality rate of 15.0%, whereas poor compliers had a mortality rate of 24.6%, so clofibrate looks promising, or so it might seem. In the placebo group, good compliers had a mortality rate of 15.1% and poor compliers had a mortality rate of 28.2%, so placebo looks even more promising than clofibrate. A comparison of groups defined in terms of compliance is a comparison of self-selected groups, which may differ in outcomes for reasons that are not effects of the treatment.

Can these two pieces of good common sense be reconciled? Can we adhere to the strict logic of the randomized experiment, adhere to comparing only groups formed by random assignment—the only randomness on which we can really count in an experiment—and yet take into account whether the prescribed drug is still in the bottle? In fact we can, as we illustrate in this article.

Robert Greevy is a doctoral candidate (E-mail: *ragreevy@wharton.upenn.edu*) and Paul R. Rosenbaum is Robert G. Putzel Professor (E-mail: *rosenbaum@stat.wharton.upenn.edu*), Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 . Jeffrey H. Silber (E-mail: *silberj@wharton.upenn.edu*) is Associate Professor of Pediatrics and Avital Cnaan (E-mail: *cnaan@email.chop.edu*) is Associate Professor of Biostatistics, Department of Pediatrics, University of Pennsylvania and The Children's Hospital of Philadelphia, Philadelphia, PA 19104. R. Greevy and P. R. Rosenbaum were supported by grant SES-0345113 from the U.S. National Science Foundation. J. H. Silber and A. Cnaan were supported by grant R01 HL-50424 from the U.S. National Heart, Lung and Blood Institute, and A. Cnaan was also supported by grant U10 CA-15488 from the U.S. National Cancer Institute.

Journal of the American Statistical Association
March 2004, Vol. 99, No. 465, Applications and Case Studies
DOI 10.1198/016214504000000025

1.2 Randomization Inference With Imperfect Compliance

An instrument manipulates a treatment without fully controlling it, but the instrument itself has no effect on the response except through its manipulation of the treatment. Instrumental variables (IVs) are commonly used in econometrics and are familiar to statisticians; for instance, Wald's (1940) estimator is essentially an IV estimator. The use of Wald's IV estimator in clinical trials with imperfect compliance is implicit in the work of Sommer and Zeger (1991) and explicit in the work of Sheiner and Rubin (1995) (see also Frangakis and Rubin 1999). Here randomization is an instrument manipulating the dose of treatment actually received. Imbens and Rubin (1997) discussed the matter from a Bayesian perspective. Angrist, Imbens, and Rubin (AIR) (1996) used the Vietnam era draft as an instrument for military service, the draft being essentially random, but with a substantial gap between the draft and actual service due to voluntary service and draft evasion. Several interesting approaches to noncompliance have been discussed by Efron and Feldman (1991), Robins and Tsiatis (1991), Robins (1994), Balke and Pearl (1997), Goetghebeur and Loeys (2002), and Henneman, van der Laan, and Hubbard (2002). These authors did not propose randomization inference for noncompliance; rather, they used sampling, modeling, and identification assumptions that go beyond the random assignment of treatments and the hypothesis being tested or inverted.

In randomization inference as developed by Fisher (1935), the null hypothesis of no treatment effect is tested without assumptions beyond the hypothesis being tested, using the fact of random assignment as the "reasoned basis for inference." Randomization tests of no effect are often inverted to yield distribution free confidence intervals and Hodges–Lehmann point estimates for additive treatment effects. (See Lehmann 1998; Cox and Reid 2000; Rosenbaum 2002a, sec. 2, for recent textbook discussions of various aspects of randomization inference.)

In the discussion of AIR (1996), Rosenbaum (1996) proposed a method of exact randomization inference with an IV, yielding exact distribution-free confidence intervals and Hodges–Lehmann nonparametric point estimates analogous to Wald's estimator. This analysis always agrees with the randomization test of no effect in the intent-to-treat analysis. They both reject or both accept the null hypothesis of no treatment effect; however, the IV test is inverted in a different way to obtain confidence statements relating dose received to magnitude of effect. The method was illustrated in two examples from labor economics by Rosenbaum (1999, 2002b) and Imbens and Rosenbaum (2004). Imbens and Rosenbaum also showed that the *only* distribution-free inferences using an IV are randomization inferences, in parallel with a result of Lehmann (1959, sec. 5.7), and that these inferences remain accurate with weak instruments for which conventional methods, such as two-stage least squares, work very poorly. In addition, with strong instruments, Imbens and Rosenbaum found in a simulation that randomization inferences have performance similar to that of two-stage least squares when all distributions are normal, but have much higher power than two-stage least squares when the response distribution is heavy-tailed.

1.3 Outline

This article applies randomization inference with an IV to the AAA randomized clinical trial with imperfect compliance. The theory is reviewed in Section 2, but certain extensions are needed to address the presence of outcome measures repeated over time, with differing periods of observation for patients entering early or late and hence differing numbers of observations for different patients. In particular, Section 2.1 recalls the statistic introduced by Wei and Lachin (1984) for clinical trials with repeated, incomplete outcome measures and uses a device of Mantel (1967) to write that statistic as a linear rank statistic, thereby simplifying use of its null distribution under randomization. That test is inverted using the IV to estimate treatment effects with incomplete compliance in Section 2.4, after the IV notation from AIR is reviewed in Section 2.3. The method is applied to the AAA study in Section 3.

2. THEORY: REVIEW AND EXTENSIONS

2.1 Testing No Effect Using the Wei–Lachin Statistic

Assume that there are I subjects, $i = 1, \ldots, I$, with n randomly assigned to treatment and the remaining $I - n$ assigned to placebo, so that all $\binom{I}{n}$ possible treatment assignments are equally probable. In the AAA study, $I = 135$ and $n = 69$. If subject i is assigned to treatment, then write $Z_i = 1$, and otherwise write $Z_i = 0$, so $n = \sum_{i=1}^{N} Z_i$. Write Y_{ik} for the outcome measurement on person i at time k, $k = 1, \ldots, K$, which may be missing due to the end of follow-up, in which case write $Y_{ik} = \varnothing$. Because patient enrollment began late in 1994 and ended in 1999, and follow-up continued until early 2001, a patient might have $K = 12$ posttreatment measures at 6-month intervals. But almost all patients were observed for a shorter period, so the later Y_{ik} are often missing; that is, for most patients, i, the response is $Y_{ik} = \varnothing$ for larger k. Missing responses created by the date of entry into the study are relatively innocuous—they do not reflect self-selection, are determined in advance solely by the date of entry into the study, and are unaffected by the treatment—however, their absence must be reflected in statistical procedures.

The null hypothesis of no treatment effect asserts that each patient i would have produced the same sequence of responses $Y_{i1}, \ldots, Y_{iK}$ whether assigned to treatment or to control. If switching the treatment for i changes i's responses, then the treatment does have at least some effect. If Y_{ik} is missing because of the date of entry of i into the study, then Y_{ik} would also have been missing had i been assigned to the alternative treatment, so the null hypothesis automatically holds for such a missing Y_{ik}. Write

$$U_{ijk} = \begin{cases} 1 & \text{if } Y_{ik} > Y_{jk} \\ -1 & \text{if } Y_{ik} < Y_{jk} \\ 0 & \text{if } Y_{ik} = Y_{jk} \text{ or} \\ & \quad \text{if either } Y_{ik} \text{ or } Y_{jk} \text{ is missing,} \end{cases} \tag{1}$$

so that $T_k = \sum_{i=1}^{I} \sum_{j=1}^{I} Z_i(1 - Z_j)U_{ijk}$ is the number of times a treated patient was observed at time k to have a higher response than a control at time k, minus the number of times that the control was observed to have the higher response at that time. Without missing data, T_k would be linearly related to the

Mann–Whitney statistic and the Wilcoxon rank sum statistic. With missing data, Wei and Lachin (1984, sec. 3) proposed using $T = \sum_{k=1}^{K} T_k$ to test for multivariate stochastic ordering of the responses in treated and control groups.

The device of Mantel (1967) permits T to be written as a linear rank statistic, which simplifies consideration of its randomization distribution. Specifically, $\sum_{i=1}^{I}\sum_{j=1}^{I} Z_i Z_j U_{ijk} = 0$ because $Z_i Z_j U_{ijk} = -Z_j Z_i U_{jik}$, so that

$$\begin{aligned} T &= \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{I} Z_i(1 - Z_j)U_{ijk} \\ &= \sum_{k=1}^{K}\sum_{i=1}^{I} Z_i \sum_{j=1}^{I} U_{ijk} \\ &= \sum_{i=1}^{I} Z_i q_i, \end{aligned} \tag{2}$$

where $q_i = \sum_{k=1}^{K}\sum_{j=1}^{I} U_{ijk}$. If the null hypothesis of no effect is true, then the q_i are fixed, not varying with different random assignments $Z_1, \ldots, Z_I$, so that the null randomization distribution of T is the distribution of the sum of n scores picked by simple random sampling without replacement from the fixed, finite population of I scores, $q_1, \ldots, q_I$, which sum to 0. For small I, the exact null randomization distribution of T may be determined by the method of Pagano and Tritchler (1983). By familiar properties of random sampling without replacement, this null randomization distribution of T has expectation 0 and variance $v = [n(I-n)/\{I(I-1)\}]\sum_{i=1}^{I} q_i^2$ (see, e.g., Rosenbaum 2002a, p. 35). Moreover, as $I \to \infty$ with n/I fixed, the null randomization distribution of the standardized deviate $T/\sqrt{v}$ converges weakly to the standard normal (see Hoeffding 1951, thm. 4).

2.2 Alternative Tests

Wei and Lachin (1984) proposed a second test that is also of use here. Using (1), write $q_{ik} = \sum_{j=1}^{I} U_{ijk}$ so that $T_k = \sum_{i=1}^{I}\sum_{j=1}^{I} Z_i(1-Z_j)U_{ijk} = \sum_{i=1}^{I} Z_i\, q_{ik}$. Under the null hypothesis of no effect, $E(T_k) = 0$ and $\mathrm{cov}(T_k, T_{k'}) = [n(I-n)/\{I(I-1)\}]\sum_{i=1}^{I} q_{ik}\, q_{ik'}$. Write $\mathbf{T} = (T_1, \ldots, T_K)^T$ and $\mathbf{\Sigma}$ for the $K \times K$ matrix of covariances $\mathrm{cov}(T_k, T_{k'})$. As an alternative to the test based on (2), Wei and Lachin (1984) suggested referring $\mathbf{T}^T\mathbf{\Sigma}^{-1}\mathbf{T}$ to tables of the chi-squared distribution on K degrees of freedom. They argued (p. 565) that the test (2) is preferable if the treatment is expected to have similar effects at all times k, whereas the alternative statistic $\mathbf{T}^T\mathbf{\Sigma}^{-1}\mathbf{T}$ is preferable if treatment is better than control at some times k and worse than control at other times (see Wei and Lachin 1984 for detailed discussion).

Another family of tests specifies a vector $\mathbf{c} = (c_1, \ldots, c_K)^T$. Under the null hypothesis of no effect, $\mathbf{c}^T\mathbf{T}$ has expectation 0 and variance $\mathbf{c}^T\mathbf{\Sigma}\mathbf{c}$, yielding the standardized deviate $\mathbf{c}^T\mathbf{T}/\sqrt{\mathbf{c}^T\mathbf{\Sigma}\mathbf{c}}$. In particular, the contrast weights $c_k = k - \frac{K+1}{2}$, or $\mathbf{c} = (-2, -1, 0, 1, 2)^T$ for $K = 5$, might be used to detect an effect that initially favors one treatment but later favors the other. Wei and Johnson (1985, sec. 4.1) proposed an adaptive choice of $\mathbf{c}$, trying to give greater weight to time periods when the effect appears larger; but unlike T in (2), this adaptive test is no longer a linear rank test with an easily used randomization distribution. Adaptive procedures may require larger sample sizes before relevant asymptotic properties apply. A related choice of $\mathbf{c}$ that is not adaptive and does yield a linear rank test is constructed using $\mathbf{\Sigma}$ as follows. Let $v_k \le n$ be the number of Y_{ik}'s observed for treated subjects at time k and let $w_k \le I - n$ be the number Y_{ik}'s observed for control subjects at time k, so that T_k makes $v_k w_k$ comparisons of actual Y_{ik}'s at time k. Write $\boldsymbol{\zeta} = (v_1 w_1, \ldots, v_K w_K)^T$. If under some alternative hypothesis, the expectation of U_{ijk} were constant for all of these $\sum v_k\, w_k$ comparisons, then the Wei and Johnson (1985, sec. 4.1) λ_k's would be the same for all times k, and their favored choice of $\mathbf{c}$ would become $\mathbf{c} = \mathbf{\Sigma}^{-1}\boldsymbol{\zeta}$. In Section 3.3 we apply this alternative test.

2.3 Effects of Assignment on Doses and Responses

This section reviews, with minor adaptations for incomplete repeated measures, the notation for experiments with instrumental variables discussed by AIR (1996). Patient i has two potential sequences of responses: (1) the sequence, say $y_{Ti1}, \ldots, y_{TiK}$, that would be seen from patient i if assigned to treatment, $Z_i = 1$, and (2) the sequence that would be seen from patient i, say $y_{Ci1}, \ldots, y_{CiK}$, if assigned to control, $Z_i = 0$. Of course, only one of the two sequences is observed from patient i, depending on the treatment assignment Z_i; that is, the observed response from i at time k is $Y_{ik} = y_{Tik}$ if i was assigned to treatment, $Z_i = 1$, and the observed response is $Y_{ik} = y_{Cik}$ if i was assigned to control, $Z_i = 0$. The effect of the treatment on patient i at time k is the unobservable quantity $y_{Tik} - y_{Cik}$. The null hypothesis of no treatment effect tested in Section 2.1 is the hypothesis that corresponding potential responses under treatment and control are the same, $y_{Tik} = y_{Cik}$, for each i, k. A constant effect, $y_{Tik} - y_{Cik} = \tau$, yields the most common of nonparametric models for two samples at each time k, namely a pair of population distributions shifted by τ. That is, the distribution of treated responses at time k—Y_{ik} with $Z_i = 1$—is the same shape and dispersion as the distribution of control responses at time k—Y_{ik} with $Z_i = 0$—but the former distribution is translated or shifted by τ. (See Neyman 1923; Welch 1937; Rubin 1974 for extensive discussion of this way to define the effects of a treatment in a randomized experiment.)

In randomization inference, as developed by Fisher (1935), quantities such as the observed response Y_{ik} that depend on the random assignment of treatments, Z_i, are random variables, whereas quantities such as the potential responses (y_{Tik}, y_{Cik}) that do not depend on treatment assignment, Z_i, are fixed features of the finite population of N subjects. In this way, randomization creates the probability distributions used for inference, and forms the "reasoned basis for inference" in Fisher's phrase. Exactly the same randomization inferences can be derived by starting with independent observations from an infinite, continuous, unknown distribution by conditioning on the order statistics, which are the complete sufficient statistics for the unknown distribution (Lehmann 1959, sec. 5.7). In this case, the model of constant effect yields the same inferences as the model of shifted distributions (see also Lehmann and Stein 1949).

Another outcome of treatment assignment is the dose, D_{ik}, of enalapril actually consumed by patient i at k. Here, too, patient i has two potential sequences of doses, $d_{Ti1}, \ldots, d_{TiK}$, if

assigned to treatment, $Z_i = 1$; and $d_{Ci1}, \ldots, d_{CiK}$, if assigned to control, $Z_i = 0$. That is, $D_{ik} = d_{Tik}$ if patient i received treatment, $Z_i = 1$, and $D_{ik} = d_{Cik}$ if patient i received placebo, $Z_i = 0$. In the AAA study, the dose of enalapril is certainly 0 for patients assigned to control, $d_{Cik} = 0$, but this is not essential, so we describe the methodology in more general terms. For instance, in a randomized trial that encouraged people to consume a readily available product, like aspirin or vitamins, as distinct from enalapril, some people assigned to control might not comply by consuming the product without being encouraged to do so.

In our analysis of the AAA study, we consider and compare several definitions of the dose of enalapril; however, each of these is scaled so that perfect compliance in the enalapril group has $d_{Ti1} = 1, \ldots, d_{TiK} = 1$ and perfect compliance in the placebo group has $d_{Ci1} = 0, \ldots, d_{CiK} = 0$. Also, $d_{Tik} = \frac{1}{2}$ signifies, for instance, consumption of half of the prescribed dose.

In contrast to the hypothesis of no treatment effect, the hypothesis

$$y_{Tik} - y_{Cik} = \beta(d_{Tik} - d_{Cik}) \tag{3}$$

says that the unobservable effect of the treatment on the response of patient i at time k is proportional to the effect on dose. If $\beta = 0$ in (3), then the treatment has no effect, $y_{Tik} = y_{Cik}$ for each i, k, as in the null hypothesis of no effect. If all patients complied perfectly, so $d_{Tik} - d_{Cik} = 1 - 0 = 1$ for every patient i and time k, then (3) is the model of a constant effect, $y_{Tik} - y_{Cik} = \beta$ yielding pairs of distributions shifted by β, as described earlier. If patient i would take no enalapril if assigned to treatment, so $d_{Tik} = 0$ for each k, then model (3) says that merely assigning the patient to a drug that he does not take has no effect, $y_{Tik} - y_{Cik} = \beta(0 - 0) = 0$, and so the patient exhibits his "no-enalapril response" whether assigned to treatment or control; that is, $Y_{ik} = y_{Cik}$ whether $Z_i = 1$ or $Z_i = 0$.

The hypothesis (3) embodies what is known as the *exclusion restriction* (AIR 1996), in which assignment to enalapril or placebo, Z_i, differentially affects the responses only indirectly by affecting the dose of enalapril consumed. The exclusion restriction is an assumption and may be false, but it is relatively plausible with a placebo. As discussed in Section 3.2, compliance appears to have been similar in the enalapril and placebo groups, so there is no indication that patients knew whether they were receiving enalapril or placebo. Without a placebo, the model (3) would say that a patient's knowledge that he or she had been assigned to receive "no enalapril" would have the same effect as being assigned to receive enalapril and consuming none of it, but this might not be true, because a patient's awareness that he or she is defying the physician's advice might, in principle, have psychological effects on cardiac health apart from the pharmacological effects of the medication. This possibility seems remote in the AAA trial, because patients assigned to placebo who refused to take it would be likely to experience similar psychological effects, if any, to patients assigned to enalapril who refused to take it. In fact, at the end of the experiment, just before families were told whether their child had received enalapril or placebo they were asked their guess, and in both groups, about half guessed correctly.

Model (3) makes no assumption about why people comply or fail to do so, and compliance may be far from random. For instance, model (3) does not preclude the possibility that the patients who would have the best responses without enalapril (i.e., the patients with the highest y_{Cik}'s) are also the patients most likely to comply by taking enalapril if assigned to it (i.e., the patients with the highest d_{Tik}'s). In such a case, good compliers in the enalapril group (i.e., patients i with $Z_i = 1$ and $D_{ik} = d_{Tik}$ near 1) are atypical; they would have had higher responses than most others even without enalapril. Alternatively, compliance could be inversely related to response without enalapril, or other factors (see Rosenbaum 2002a, sec. 5.4.4, for a small, artificial example illustrating this issue).

2.4 Randomization Inference With an Instrumental Variable

Randomization inference with an instrumental variable is straightforward (Rosenbaum 1996, 1999, sec. 5, 2002a,b; Imbens and Rosenbaum 2004). Under the model (3),

$$y_{Tik} - \beta d_{Tik} = y_{Cik} - \beta d_{Cik} = a_{ik}, \quad \text{say,} \tag{4}$$

is fixed, not varying with Z_i. Consider testing $H_0 : \beta = \beta_0$ in (3). Again, in the AAA study $d_{Cik} = 0$, because controls had no access to enalapril, so in this case $a_{ik} = y_{Cik}$; however, this is not essential to the argument, which is presented in more general terms. The observed response Y_{ik} and the observed dose D_{ik} are random variables, changing with treatment assignment, Z_i, but if the null hypothesis $H_0 : \beta = \beta_0$ were true, then $Y_{ik} - \beta_0 D_{ik} = a_{ik}$ is the fixed quantity in (4), which is actually y_{Cik} in the AAA study. Hence $H_0 : \beta = \beta_0$ can be tested by applying the test statistic in Section 2.1 to the $Y_{ik} - \beta_0 D_{ik}$. The set of values of β_0 not rejected at the .05 level is a 95% confidence set for β (e.g., Lehmann 1959, sec. 3.5). The value of $\widehat{\beta}$ that equates (as closely as possible) the test statistic T in (2) computed from $Y_{ik} - \beta_0 D_{ik}$ to its null expectation, namely 0, is the Hodges–Lehmann point estimate (Hodges and Lehmann 1963) of β.

Instead of reporting a 95% confidence interval, a point estimate, and a standard error, Mosteller and Tukey (1977) suggested reporting a 95% confidence interval, a point estimate, and a 2/3 confidence interval. If the point estimate were asymptotically normal, then the 2/3 confidence interval would be close to the point estimate plus or minus its standard error. Mosteller and Tukey's suggested approach has several advantages. If the estimate is not asymptotically normal, for instance (e.g., if its distribution is long tailed or highly skewed), then the estimate may not have a finite standard error, or else the estimate plus or minus a standard error may have little meaning. In contrast, a 2/3 confidence interval is meaningful without asymptotic normality, meaningful for estimators with skewed distributions, and meaningful in small samples without reference to asymptotic theory. Like the standard error, the 2/3 confidence interval provides some guidance in thinking about a point estimate, because all values of the parameter in the 2/3 confidence interval are fully compatible with the observed data. With weak instruments, the distribution of IV estimates is often poorly approximated by a normal distribution (Nelson and Startz 1990; Maddala and Jeong 1992), and the only exact, nonparametric confidence intervals are derived from permutation

or randomization methods (Imbens and Rosenbaum 2004). For this reason, with IVs it is wise to follow Mosteller and Tukey's suggestion, replacing standard errors by 2/3 confidence intervals.

2.5 Properties of the Inference: Agrees With Intent-to-Treat; Consistent

It is important to note that the test just described always agrees with the intent-to-treat analysis about whether the null hypothesis of no effect is plausible. Specifically, there is no effect in (3) if $\beta = 0$, and the test of $H_0 : \beta = 0$ calculates the Wei–Lachin statistic from the $Y_{ik} - 0D_{ik} = Y_{ik}$ and compares the statistic with its randomization distribution, which is of course exactly the corresponding test of no effect in the intent-to-treat analysis that ignores dose information.

In the AAA study there are certain simplifications that aid in understanding the behavior of the test when the null hypothesis is false, say $\beta > \beta_0$. Specifically, consider what happens as $I \to \infty$ with $n/I \to 1/2$, and suppose that at least some enalapril patients take at least some of their medication, or, precisely, suppose that a fraction $\xi > 0$ of the patients takes at least a fraction $\eta > 0$ of their drug at each time k for all sufficiently large I. Then the test is consistent against $H_A : \beta > \beta_0$; that is, the probability of rejection tends to 1 as $I \to \infty$. To see this, note that in the AAA study, controls received no enalapril, $d_{Cik} = 0$, so that $a_{ik} = y_{Cik}$ in (4), and $D_{ik} = 0$ for controls with $Z_i = 0$. Therefore, the adjusted response is

$$Y_{ik} - \beta_0 D_{ik} = Y_{ik} - \beta D_{ik} + (\beta - \beta_0) D_{ik} = a_{ik} + (\beta - \beta_0) D_{ik},$$

which for a control with $Z_i = 0$ equals y_{Cik} and for a treated patient with $Z_i = 1$ equals $y_{Cik} + (\beta - \beta_0) d_{Tik} \geq y_{Cik}$ with strict inequality under H_A for at least a fraction, $\xi > 0$, of the patients with $d_{Tik} > \eta$. Under H_A, the limiting distribution of $Y_{ik} - \beta_0 D_{ik}$ for treated patients, $Z_i = 1$, will be strictly stochastically larger than the limiting distribution of $Y_{ik} - \beta_0 D_{ik}$ for controls, $Z_i = 0$, and Wei and Lachin's sum of rank sum statistics will have power tending to 1.

2.6 An Alternative Effect Model Leading to the Same Inferences

This section describes an alternative view of the proposed procedure. In the alternative model, (3) does not hold exactly, but rather holds with stochastic errors, but this alternative model yields the same permutation inferences. In the current section only, replace (3) with $y_{Tik} = \mu + \beta d_{Tik} + \varepsilon_{Tik}$, $y_{Cik} = \mu + \beta d_{Cik} + \varepsilon_{Cik}$. Write $\varepsilon_{Ti} = (\varepsilon_{Ti1}, \ldots, \varepsilon_{TiK})$ and $\varepsilon_{Ci} = (\varepsilon_{Ci1}, \ldots, \varepsilon_{CiK})$ for the two possible sequences of errors for subject i, and assume that (a) the $(\varepsilon_{Ti}, \varepsilon_{Ci})$ for different subjects i are mutually independent, (b) the errors $(\varepsilon_{Ti}, \varepsilon_{Ci})$ are independent of the random assignments Z_i to enalapril or placebo, and (c) the pair of error vectors for person i is exchangeable in the sense that $\Pr\{(\varepsilon_{Ti}, \varepsilon_{Ci}) = (\mathbf{a}, \mathbf{b})\} = \Pr\{(\varepsilon_{Ti}, \varepsilon_{Ci}) = (\mathbf{b}, \mathbf{a})\}$ for each $(\mathbf{a}, \mathbf{b})$. In this case, the treatment effect, $y_{Tik} - y_{Cik} = \beta(d_{Tik} - d_{Cik}) + (\varepsilon_{Tik} - \varepsilon_{Cik})$, is symmetrically distributed about, but not equal to, $\beta(d_{Tik} - d_{Cik})$. If $H_0 : \beta = \beta_0$ is true, then calculate $Y_{ik} - \beta_0 D_{ik} = \mu + E_{ik} = a_{ik}$, say, where the $E_{ik} = Z_i \varepsilon_{Tik} + (1 - Z_i)\varepsilon_{Cik}$ are independent of the assigned treatments, Z_i, because $(\varepsilon_{Tik}, \varepsilon_{Cik})$ is exchangeable and independent of Z_i. If $H_0 : \beta = \beta_0$ is true, then the conditional distribution of the Z_i's given the a_{ik}'s is still the randomization distribution, because the Z_i's are independent of the a_{ik}'s, so viewing the a_{ik} as fixed in Section 2.4 is the same as conditioning on their observed values, resulting in an appropriate test (see Rosenbaum 2002b, rejoinder secs. 3 and 4; Imbens and Rosenbaum 2004 for further discussion along these lines).

3. INSTRUMENTAL VARIABLES ANALYSIS OF THE AAA STUDY

3.1 Structure of the AAA Study

In the AAA study, 135 children were randomized, with 69 receiving enalapril and 66 receiving placebo, at 4 institutions. Before treatment, the two groups did not differ significantly with respect to age, weight, body mass index, systolic or diastolic blood pressure, and various cardiac measures, including the two variables that were measured later as outcomes, maximum cardiac index and left-ventricle end-systolic wall stress (see Silber et al. 2004, table 2).

Cardiac performance testing was done for each patient at baseline and then at 6-month intervals after treatment. Patients provided varied numbers of performance tests because of varied dates of entry into the study. We have information about consumption of placebo, and although it does not enter into the primary analyses, we do use this information for some purposes.

The intent-to-treat analysis provided no evidence of effect on the primary outcome, the change in the maximum cardiac index. Of course, this is an important clinical finding; however, because our IV analysis always agrees with the intent-to-treat analysis about rejection or acceptance of the null hypothesis of no effect, it has little to add about the maximum cardiac index, for which the null hypothesis of no effect is plausible. For this reason, we focus attention here on another outcome, change in left ventricular end-systolic wall stress, recorded in g/cm^2. The analysis presented here is intended solely to illustrate statistical methodology, not to reach clinical conclusions, so we explore statistical issues in detail and present only aspects of the clinical results selected for their relevance to the statistical issues. (For all substantive clinical findings, see Silber et al. 2004.)

3.2 Compliance

At 3-month intervals, patients were asked about prescribed pills taken or missed in the previous week. For patient i at month m, the variable P_{im} records a 1 if the prescribed dose was consumed, 0 if no dose was consumed, 1/2 if half the prescribed dose was consumed, and so on. The outcome, the change in wall stress from baseline, is available only at 6-month intervals, not 3-month intervals. To make use of all available reports on pills consumed, we applied exponential smoothing (Cox 1961), in which the smoothed pill consumption, P^s_{im}, for patient i at month m is a weighted average of the *reported* pill consumption, P_{im}, at month m and the *smoothed* consumption, $P^s_{i,m-3}$, at the previous reporting period, $P^s_{im} = (1 - \lambda)P_{im} + \lambda P^s_{i,m-3}$, with $P^s_{i0} = 1$, and with $P^s_{im} = P^s_{i,m-3}$ in the rare instances when P_{im} was missing. In most analyses, we used $\lambda = \frac{1}{2}$, so $P^s_{im} = \frac{1}{2}P_{im} + \frac{1}{2}P^s_{i,m-3}$, where $P^s_{i,m-3}$ is defined similarly. So on expanding this expression, one finds that P^s_{im} is the simple average of the

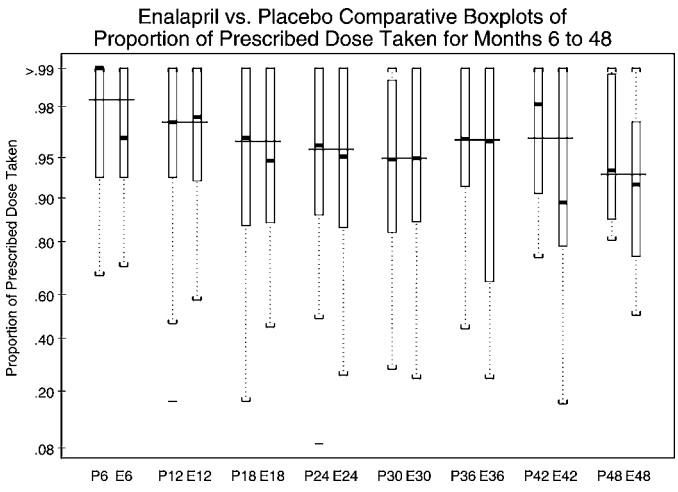

Figure 1. Compliance, P^s_{im}, in the Enalapril (E) and Placebo (P) Groups for the First 48 Months.

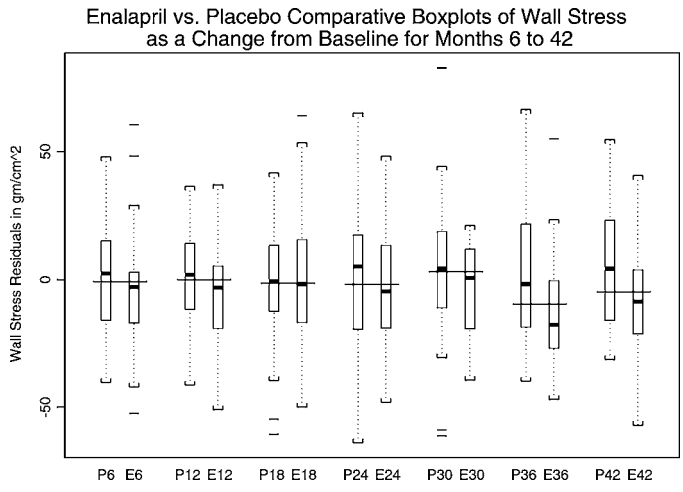

Figure 2. Change in Wall Stress From Baseline, Y_{ik}.

currently reported value P_{im} and a weighted average $P^s_{i,m-3}$ of all previous values giving greatest weight to the recent past, $P^s_{im} = \frac{1}{2}P_{im} + \frac{1}{4}P_{i,m-3} + \frac{1}{8}P_{i,m-6} + \cdots$. We varied both λ and the missing-value rule, but the results changed only slightly.

Figure 1 plots, on a logit scale, the smoothed pill consumptions, P^s_{im}, at 6-month intervals for the first 48 months. The short horizontal lines in the figures are at the pooled median at time k, combining the enalapril and placebo groups. Compliance was typically quite good; in all of the boxplots, the lower quartile of P^s_{im} is above 80% and the upper quartile is 100%. Compliance was somewhat better at months 6 and 12 than at later months. There are some instances of poor compliance, say $\leq 50\%$. When the Wei–Lachin statistic T is applied to the pill consumptions, P^s_{im}, the difference is not significant, with standardized deviate $-.57$ and p value .57, so there is no indication that compliance in the enalapril and placebo groups is dissimilar.

Recall that the outcome is available at 6-month intervals, so the kth time period corresponds with month $m = 6k$. The dose of enalapril at time k is defined as $D_{ik} = P^s_{i,6k}$ for patients in the enalapril group and as $D_{ik} = 0$ for patients in the placebo group.

3.3 Effects on Change in Wall Stress

Figure 2 displays the change in wall stress from baseline for enalapril and placebo groups for 6–42 months. The hope was that enalapril would reduce wall stress when compared to placebo, and at every time k, the medians do appear slightly lower in the enalapril group. The boxplots call attention to a few extreme values.

Table 1 reports two analyses, the intent-to-treat analysis and the IV analysis. The IV analysis uses the hypothesis (3) that effect is proportional to dose and draws inferences about β, which is the effect of enalapril with full compliance or dose $D_{ik} = 1$. The intent-to-treat analysis makes no use of the compliance information, D_{ik}, and simply compares the enalapril group as a whole to the placebo group as a whole. The intent-to-treat analysis with an additive effect or shift in location is the same as the hypothesis (3) but with all doses equal to 1 in the enalapril group and all doses equal to 0 in the placebo group; it estimates the effect of encouraging someone to take enalapril rather than placebo as distinct from the effect of actually taking enalapril. Table 1 gives the two-sided significance level for testing no effect, the Hodges–Lehmann point estimate of effect, and the 95%, 90%, and 2/3 confidence intervals. Recall from Section 2.4 that the 2/3 confidence interval is Mosteller and Tukey's (1977) suggested replacement for a point estimate $\pm$ a standard error.

Table 1 shows that, as always, the significance level for testing the hypothesis of no effect is exactly the same for the intent-to-treat and IV analyses. The point estimate of the effect of being assigned to the enalapril group rather than to the control group is -6.57 g/cm^2, whereas the point estimate of the effect of full compliance with the assigned treatment is -7.11 g/cm^2. Comparing the two Hodges–Lehmann point estimates, the effect of a full dose of enalapril is estimated to be about 8% larger than the effect of encouragement to take enalapril, that is, $-7.11/(-6.57) - 1 = 8.2\%$.

Table 1 was built by inverting the Wei–Lachin test under the hypothesis (3), as described in Section 2.4. Specifically, the hypothesis $H_0 : \beta = \beta_0$ was tested by subtracting the hypothesized effect, $\beta_0 D_{ik}$, from the observed change in wall stress, Y_{ik}, and applying the Wei–Lachin statistic to the adjusted responses, $Y_{ik} - \beta_0 D_{ik}$, which would be free of the enalapril effect if the null hypothesis were true. The confidence interval is the set of hypotheses not rejected by the test, and the Hodges–Lehmann point estimate is the value of β_0 that equates the Wei–Lachin

Table 1. Intent-to-Treat and IV Analyses of Wall Stress

Intent-to-treat			
p value for no effect		.0449	
Hodges–Lehmann estimate		−6.57	
95% confidence interval	−12.60		−.10
90% confidence interval	−11.70		−1.09
2/3 confidence interval	−8.00		−5.07
IV			
p value for no effect		.0449	
Hodges–Lehmann estimate		−7.11	
95% confidence interval	−13.82		−.11
90% confidence interval	−12.70		−1.14
2/3 confidence interval	−8.81		−5.65

statistic as nearly as possible to its null expectation. Both the intent-to-treat analysis and the IV analysis in Table 1 use this same logic, and in both the inference is based solely on the random assignment of treatments and the hypothesis that is being tentatively assumed for the purpose of testing it. The difference between the two analyses is in the family of hypotheses being tested. The intent-to-treat analysis works with hypotheses asserting that enalapril is equally effective whether the patient consumes the assigned dose or not. The IV analysis works with hypotheses asserting that the effect of enalapril is proportional to dose, so that 0 dose has no effect. Because both analyses use randomization as the basis for inference using the same test statistic, a preference for one analysis over the other represents a preference for one family of hypotheses over the other.

As discussed in Section 2.2, an alternative to the test statistic (2) is the test statistic $T^* = \mathbf{c}^T\mathbf{T}$ with $\mathbf{c} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\zeta}$, which has certain optimal properties under certain assumptions. How does T^* perform in the AAA study? As different hypotheses $H_0: \beta = \beta_0$ are tested, the q_{ik} change; so $\boldsymbol{\Sigma}$ changes, and hence the weights $\mathbf{c}$ also change. Using T^* as the test statistic, (a) the significance level for testing no effect is .026, compared with .045 in Table 1; (b) the 95% confidence interval for β is $[-14.02, -.82]$, which is about 4% shorter than the corresponding interval $[-13.83, -.11]$ in Table 1 and is slightly further from 0; and (c) the point estimate is $\widehat{\beta} = -7.50$, as opposed to $\widehat{\beta} = -7.11$ in Table 1. The weights, $\mathbf{c} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\zeta}$ exhibit unanticipated behavior, however. For instance, when testing no effect, $H_0: \beta = 0$, most c_k's are positive, but c_9, c_{11}, and c_{12} are quite negative and c_6 is slightly negative, whereas c_{10} is positive and four times larger than the next-largest positive weight. Moreover, c_3 and c_6 are about one-tenth the size of most other c's. This means that T^* counts declines in wall stress at times $k = 3$ and $k = 6$ as unimportant; declines at times $k = 9$, 11, and 12 as unfavorable; and declines at other times as favorable, especially favorable at time $k = 10$. Because it was hoped that enalapril would reduce wall stress at all times, these weights, $\mathbf{c} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\zeta}$, did not seem to match the goals of the trial.

3.4 Informal Checking of the Model

The model (3), which defines the parameter β, says that treatment effect at time k is proportional to dose effect at time k. If the model were true, then treatment assignment Z_i would be independent of $Y_{ik} - \beta D_{ik} = a_{ik}$, which in the AAA study is also equal to the potential response under placebo, y_{Cik}. Although β is unknown, we can look at $Y_{ik} - \widehat{\beta} D_{ik}$.

The model (3) could be false in a variety of ways, including the following. First, the difference between the enalapril and placebo groups might not be adequately described as a function of dose D_{ik} alone; that is, the exclusion restriction might be false. Second, even if a function of dose described the effect, it might not be a linear function, βD_{ik}. Third, even if a linear function did describe the effect, it might require different linear functions, say $\beta_k D_{ik}$, at different times k.

Figure 3 presents pairs of boxplots of $Y_{ik} - \widehat{\beta} D_{ik}$ for enalapril ($Z_i = 1$) and placebo ($Z_i = 0$) at 6-month intervals (k) for 6–42 months, after which the data become rather thin. That is, Figure 3 describes the responses in both treatment groups at various times after removing the estimated effect of enalapril, $\widehat{\beta} D_{ik}$. If the model (3) were true, then $Y_{ik} - \beta D_{ik}$ would be independent of Z_i, so a pair of boxplots of $Y_{ik} - \beta D_{ik}$ at a time k would not exhibit systematic differences. In fact, Figure 3 does not reveal any highly systematic differences between the pair of boxplots at each k. At each time k, the two boxplots look relatively similar. At some times k, the enalapril group is a little higher or more dispersed than the placebo group, but this reverses at other times, without a definite pattern.

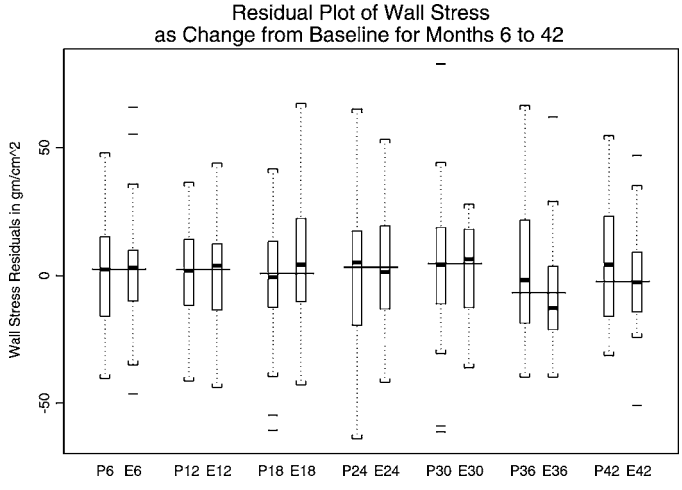

Figure 3. Residuals, $Y_{ik} - \widehat{\beta} D_{ik}$, by Assigned Group.

Figure 3 exhibits two interesting features that are not consequences of the model. First, the medians of $Y_{ik} - \widehat{\beta} D_{ik}$ hover in the vicinity of 0. Model (3) provides no reason to anticipate this; rather, in this particular study Figure 3 suggests that changes in wall stress from baseline would have a median of about 0 aside from effects, if any, of enalapril. Second, the boxplots of $Y_{ik} - \widehat{\beta} D_{ik}$ do not appear to be changing in a highly systematic way over time k. Again, this is not implied by the model (3); that model implies only that the enalapril–placebo pair of boxplots at each time k will be similar, not that boxplots for different times k will be similar.

3.5 Testing the Model

The small differences between pairs of boxplots in Figure 3 appear erratic and innocuous, but the eye is not always the best judge. Also, Figure 3 acted as if $\widehat{\beta}$ were the true value of β, and this is hardly likely. In this section, we take a slightly more formal approach to judge whether the model (3) is plausible, while continuing to use randomization tests. Roughly speaking, we look at the most plausible values of β_0, namely those in the 2/3 confidence interval in Table 1, and test at level .05 whether the model is plausible for these β_0. In a somewhat related but different approach, a p value maximized over a confidence set was used in a different way by Berger and Boos (1994).

Suppose that we have two tests of a specific model and a parameter value, for instance, two tests of the conjunction of the family of models (3) together with the specific hypothesis $H_0: \beta = \beta_0$. The first test rejects a true model/parameter $100\alpha_1\%$ of the time, and it is designed to be sensitive to the value of the parameter assuming that the model is true. The second test rejects a true model/parameter $100\alpha_2\%$ of the time

and is designed to be sensitive to certain ways in which the model may be false; specifically, it is a consistent test of the model against a certain class $\mathcal{C}$ of alternatives to the model. If we test $H_0: \beta = \beta_0$ twice, once using each test, then by the Bonferroni inequality, then the chance of falsely rejecting the true β_0 is at most $\alpha_1 + \alpha_2$, and the set of β_0 not rejected in this way covers the true β_0 in at least $100(1 - \alpha_1 - \alpha_2)\%$ of experiments. If $\alpha_1 = 1/3$ and $\alpha_2 = .05$, then β_0 is inside the two-step confidence interval if and only if it is inside the 2/3 confidence interval based on the first test, and not rejected as implausible at the .05 level by the second test. If the model is true, then the two-step interval covers the true β in at least $100(1 - \frac{1}{3} - \frac{1}{20})\% = 61\frac{2}{3}\% \doteq 3/5$ of experiments. Although this is in principle a shorter interval with lower coverage, the values β_0 in the interval are highly plausible, and for these values the model itself is not implausible. If the model is false and one of the alternatives in $\mathcal{C}$ is true instead, then the probability that the two-stage interval is empty tends to 1 as the sample size increases. These are the only formal operating characteristics of the two-stage testing procedure.

The specific procedure is to (1) build the 2/3 confidence interval for β under model (3) by testing each hypothesis $H_0: \beta = \beta_0$ using T in (2) applied to $Y_{ik} - \beta_0 D_{ik}$, then (2) test at the .05 level the model (3) and each β_0 in this 2/3 confidence interval using $\mathbf{T}^T \boldsymbol{\Sigma}^{-1} \mathbf{T}$ in Section 2.2 applied to $Y_{ik} - \beta_0 D_{ik}$ (see the Appendix for additional discussion). We applied this procedure using the data for months 6–42 depicted in Figure 3, that is, 7 time periods, $k = 1, \ldots, 7$, because each of the 7 Mann–Whitney–Wilcoxon statistics, T_k, was then based on at least 20 enalapril and 20 placebo patients. Hence, $\mathbf{T}^T \boldsymbol{\Sigma}^{-1} \mathbf{T}$ is compared to a chi-squared distribution on 7 degrees of freedom, with a upper 5% point of 14.07. In fact, for all β_0 in the 2/3 confidence interval in the IV analysis in Table 1, $\mathbf{T}^T \boldsymbol{\Sigma}^{-1} \mathbf{T}$ is less than 3.91. Formally, the 2/3 confidence interval for β in Table 1 is also a 3/5 confidence interval testing both parameter value and model specification. Informally, there is not the slightest sign that the minor ups and downs in adjacent pairs of enalapril/placebo boxplots in Figure 3 represent systematic patterns; for $\widehat{\beta}$ in the plot and for every β_0 in the 2/3 confidence interval, the fluctuations among the T_k are relatively small compared to the fluctuations expected by chance, that is, from random assignment of treatments when (3) is correct.

We repeated this approach with a different second test—a test based on the linear contrast, $\mathbf{c}^T \mathbf{T} / \sqrt{\mathbf{c}^T \boldsymbol{\Sigma} \mathbf{c}}$, with $\mathbf{c} = (-3, -2, -1, 0, 1, 2, 3)^T$—to look for a linear trend in the Mann–Whitney–Wilcoxon statistics, T_k, $k = 1, \ldots, 7$. Again, we calculated $|\mathbf{c}^T \mathbf{T}| / \sqrt{\mathbf{c}^T \boldsymbol{\Sigma} \mathbf{c}}$ from $Y_{ik} - \beta_0 D_{ik}$ for every β_0 in the 2/3 confidence interval, and the maximum standardized normal deviate was .63. In short, there is no compelling evidence against the model (3) of effect proportional to dose for the highly plausible values of β.

3.6 Comparison With a Per-Protocol Analysis

The more common, but (we believe) less satisfactory analysis is the per-protocol analysis, which excludes noncompliers. This analysis is not justified by randomization, because the excluded subjects were excluded by their own actions, not at random. Moreover, compliance is a matter of degree, but the per-protocol analysis classifies patients into compliers and noncompliers. For illustration here, we defined a complier as a patient who took $\geq 75\%$ of the prescribed medication in total over all observation periods. There were 62/69 compliers in the enalapril group and 60/66 compliers in the placebo group. If the Wei–Lachin test (2) is applied to the change in wall stress for the compliers, the two-sided p value is .11, in contrast to .045 in Table 1 for all subjects, and the point estimate of a constant effect is -5.40, in contrast to -6.57 in Table 1. The groups of noncompliers are too small to permit stable estimates. So the per-protocol analysis did not quite reject the null hypothesis of no effect, whereas the intent-to-treat analysis just barely rejected it, and of these two tests, only the intent-to-treat analysis is justified by randomization. In contrast, the intent-to-treat analysis and the IV analysis in Table 1 agree exactly about the null hypothesis of no effect, and both are randomization tests.

4. SUMMARY

Our analysis used randomization as the "reasoned basis for inference," in Fisher's (1935) phrase, and yet it addressed imperfect compliance with the study protocol. That is, our hypothesis test assumes nothing beyond the use of randomization in assigning subjects to enalapril or control, and the null hypothesis being tested, namely (3) for a specific $H_0: \beta = \beta_0$. No assumption was made that compliance was random; quite possibly compliance was severely biased. This test always agrees exactly with the randomization test of no effect in an intent-to-treat analysis—the significance levels are identical. Confidence intervals were obtained by inverting the test. In short, one can adhere to the strict logic of the randomized experiment, yet acknowledge that a drug that remains in the bottle will be without biological effect.

APPENDIX: REMARKS ON THE REINFORCED CONFIDENCE INTERVALS

We describe the two-step confidence interval in Section 3.5 in more explicit terms. Consider testing $H_0: \beta = \beta_0$ using the two test procedures in Section 3.5, and let $\mathbf{T}_{\beta_0}$ be the corresponding statistic and let $\boldsymbol{\Sigma}_{\beta_0}$ be its null covariance matrix, as defined in Section 2.2. Using the reasoning associated with Scheffé's multiple comparison procedure, it follows that

$$\mathbf{T}_{\beta_0}^T \boldsymbol{\Sigma}_{\beta_0}^{-1} \mathbf{T}_{\beta_0} \leq a^2 \quad \text{if and only if}$$
$$-a\sqrt{\mathbf{c}^T \boldsymbol{\Sigma}_{\beta_0} \mathbf{c}} \leq \mathbf{c}^T \mathbf{T}_{\beta_0} \leq a\sqrt{\mathbf{c}^T \boldsymbol{\Sigma}_{\beta_0} \mathbf{c}} \quad \text{for all } \mathbf{c} \quad \text{(A.1)}$$

(see, e.g., Hsu 1996, lemma B.1.1, p. 231, and sec. 7.5). The 2/3 confidence interval includes β_0 if the second part of (A.1) is true with $\mathbf{c} = \mathbf{1} = (1, 1, \ldots, 1)^T$ and $a = .96$ or approximately $a \doteq 1$. There is rejection of $H_0: \beta = \beta_0$ at the second stage at the .05 level if the first part of (A.1) is false with a^2 equal to the upper .05 percentage point of chi-squared with K degrees of freedom, or $a = \sqrt{14.07} = 3.75$ in Section 3.5. In other words, $H_0: \beta = \beta_0$ is not rejected in the two-stage procedure in Section 3.5 if $|\mathbf{1}^T \mathbf{T}_{\beta_0}| / \sqrt{\mathbf{1}^T \boldsymbol{\Sigma}_{\beta_0} \mathbf{1}} \leq .96$ and $|\mathbf{c}^T \mathbf{T}_{\beta_0}| / \sqrt{\mathbf{c}^T \boldsymbol{\Sigma}_{\beta_0} \mathbf{c}} \leq 3.75$ for every $\mathbf{c}$.

[Received January 2003. Revised December 2003.]

REFERENCES

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455.

Balke, A., and Pearl, J. (1997), "Bounds on Treatment Effects From Studies With Imperfect Compliance," *Journal of the American Statistical Association*, 92, 1171–1176.

Berger, R. L., and Boos, D. D. (1994), "*P* Values Maximized Over a Confidence Set for the Nuisance Parameter," *Journal of the American Statistical Association*, 89, 1012–1016.

Cox, D. R. (1961), "Prediction by Exponentially Weighted Moving Averages and Related Methods," *Journal of the Royal Statistical Society*, Ser. B, 23, 414–422.

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

Efron, B., and Feldman, D. (1991), "Compliance as an Explanatory Variable in Clinical Trials," *Journal of the American Statistical Association*, 86, 9–17.

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

Frangakis, C. E., and Rubin, D. B. (1999), "Addressing Complications of Intention-to-Treat Analysis in the Combined Presence of All-or-None Treatment Noncompliance and Subsequent Missing Outcomes," *Biometrika*, 86, 365–379.

Goetghebeur, E., and Loeys, T. (2002), "Beyond Intent to Treat," *Epidemiologic Reviews*, 24, 85–90.

Henneman, T. A., van der Laan, M. J., and Hubbard, A. E. (2002), "Estimating Causal Parameters in Marginal Structural Models With Unmeasured Confounders Using Instrumental Variables," Working Paper 104, University of California at Berkeley, Division of Biostatistics, available at *http://www.bepress.com/ucbbiostat/paper104*.

Hodges, J. L., and Lehmann, E. L. (1963), "Estimates of Location Based on Ranks," *The Annals of Mathematical Statistics*, 34, 598–611.

Hoeffding, W. (1951), "A Combinatorial Central Limit Theorem," *The Annals of Mathematical Statistics*, 22, 558–566.

Hsu, J. C. (1996), *Multiple Comparisons: Theory and Methods*, London: Chapman & Hall.

Imbens, G. W., and Rosenbaum, P. R. (2004), "Robust, Accurate Confidence Intervals With a Weak Instrument: Quarter of Birth and Education," *Journal of the Royal Statistical Society*, Ser. A, 167, to appear.

Imbens, G. W., and Rubin, D. B. (1997), "Bayesian Inference for Causal Effects in Randomized Experiments With Noncompliance," *The Annals of Statistics*, 25, 305–327.

Lehmann, E. L. (1959). *Testing Statistical Hypotheses*, New York: Wiley.

——— (1998), *Nonparametrics*, Upper Saddle River, NJ: Prentice-Hall.

Lehmann, E., and Stein, C. (1949), "On the Theory of Some Nonparametric Hypotheses," *The Annals of Mathematical Statistics*, 20, 28–45.

Maddala, G. S., and Jeong, J. (1992), "On the Exact Small-Sample Distribution of the Instrumental Variable Estimator," *Econometrica*, 60, 181–183.

Mantel, N. (1967), "Ranking Procedures for Arbitrarily Restricted Observations," *Biometrics*, 23, 65–78.

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.

Moses, L. E. (1965), "Confidence Limits From Rank Tests," *Technometrics*, 7, 257–260.

Mosteller, F., and Tukey, J. (1977), *Data Analysis and Regression*, Waltham, MA: Addison-Wesley.

Nelson, C. R., and Startz, R. (1990), "Some Further Results on the Exact Small-Sample Properties of the Instrumental Variable Estimator," *Econometrica*, 58, 967–976.

Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles," *Roczniki Nauk Roiniczych*, Tom X, Sec. 9, pp. 1–51 (in Polish); reprinted with discussion in *Statistical Science* (1990), 5, 463–480.

Pagano, M., and Tritchler, D. (1983), "Obtaining Permutation Distributions in Polynomial Time," *Journal of the American Statistical Association*, 78, 435–440.

Robins, J. M. (1994), "Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models," *Communications in Statistics, Part A—Theory and Methods*, 23, 2379–2412.

Robins, J. M., and Tsiatis, A. A. (1991), "Correcting for Non-Compliance in Randomized Trials Using Rank Preserving Structural Failure Time Models," *Communications in Statistics, Part A—Theory and Methods*, 20, 2609–2631.

Rosenbaum, P. R. (1996), Comment on "Identification of Causal Effects Using Instrumental Variables," by J. D. Angrist, G. W. Imbens, and D. B. Rubin, *Journal of the American Statistical Association*, 91, 465–468.

——— (1999), "Using Combined Quantile Averages in Matched Observational Studies," *Applied Statistics*, 48, 63–78.

——— (2002a), *Observational Studies* (2nd ed.), New York: Springer-Verlag.

——— (2002b), "Covariance Adjustment in Randomized Experiments and Observational Studies" (with discussion), *Statistical Science*, 17, 286–327.

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.

Sheiner, L. B., and Rubin, D. B. (1995), "Intention-to-Treat Analysis and the Goals of Clinical Trials," *Clinical Pharmacology and Therapeutics*, 57, 6–15.

Silber, J. H., Cnaan, A., Clark, B. J., Paridon, S. M., Chin, A. J. et al. (2001), "Design and Baseline Characteristics for the ACE-Inhibitor After Anthracycline (AAA) Study of Cardiac Dysfunction in Pediatric Oncology Long-Term Survivors," *American Heart Journal*, 142, 577–585.

Silber, J. H., Cnaan, A., Clark, B. J., Paridon, S. M., Chin, A. J. et al. (2004), "Enalapril to Prevent Cardiac Function Decline in Long-Term Survivors of Pediatric Cancer Exposed to Anthracyclines," *Journal of Clinical Oncology*, 5, 820–828.

Sommer, A., and Zeger, S. L. (1991), "On Estimating Efficacy From Clinical Trials," *Statistics in Medicine*, 10, 45–52.

Wald, A. (1940), "The Fitting of Straight Lines if Both Variables Are Subject to Error," *The Annals of Mathematical Statistics*, 11, 284–300.

Wei, L. J., and Johnson, W. E. (1985), "Combining Dependent Tests With Incomplete Repeated Measurements," *Biometrika*, 72, 359–364.

Wei, L. J., and Lachin, J. M. (1984), "Two-Sample Asymptotically Distribution-Free Tests for Incomplete Multivariate Observations," *Journal of the American Statistical Association*, 79, 653–661.

Welch, B. L. (1937), "On the z-Test in Randomized Blocks and Latin Squares," *Biometrika*, 29, 21–52.