

Often, the protocol for a randomized controlled trial states that if a patient's health deteriorates substantially in certain specific ways, then the study treatments will no longer be considered adequate medical care, and the patient will be taken off the study treatments and given the needed care. Aberrant responses of this sort present special problems of interpretation. The few such patients may have received only limited exposure to the study treatments, but also may have experienced some of the worst outcomes observed in the trial. The most basic question is whether one of the treatments tends to cause such aberrant responses, quite apart from any other beneficial or harmful effects that treatment may have. Common practice is to reduce aberrant responses to binary events and apply Fisher's exact test for a 2×2 table, but this approach discards the available information about the magnitude of the aberration. We propose a new randomization test for this problem that uses information about both the number of aberrations and their magnitude. In extreme situations, the new test reduces to either Fisher's exact test or Wilcoxon's rank sum test, but in common situations, the new test has features of both methods. Importantly, this exact test reacts only to aberrant effects; it is completely unaffected by treatment effects that do not produce aberrations. When aberrant responses have several aspects, we examine the consequences of selecting, after the fact, the one aspect exhibiting the greatest aberration, and applying the test to that aspect. For a single response variable, a confidence interval is proposed for an additive aberrant effect; the interval is constructed by inverting the test, and a multivariate extension is briefly mentioned. We illustrate the use of the test in the ACE-Inhibitor After Anthracycline (AAA) randomized trial, which was aimed at preserving cardiac function in children treated for cancer. Seven of 135 children were removed from the trial due to substantial cardiac declines; 6 of these were on placebo, 1 of whom subsequently died.

KEY WORDS: Causal effect; Clinical trial; Multiplicity in testing; Permutation test; Randomization test; Randomized experiment.

1. ABERRANT EFFECTS: INTRODUCTION AND AN EXAMPLE

1.1 Introduction: Studying Preventive Care Amid Life-Threatening Events

The design protocol for many randomized clinical trials states that if a patient's condition severely deteriorates in certain specified ways, so that treatments under study no longer constitute appropriate or adequate care, then the patient will be removed from the study's treatments and given the care the patient is deemed to need. For instance, the example presented in Sections 1.2 and 2.4 concerns the ACE-Inhibitor After Anthracycline (AAA) randomized trial, in which patients received either a double blind placebo or the drug enalapril as a mild, preventive measure intended to maintain cardiac function after cancer chemotherapy with anthracyclines (Silber et al. 2001, 2004; Greevy, Silber, Cnaan, and Rosenbaum 2004). The protocol stated that if a patient's cardiac function declined substantially in certain specified ways, then mild preventive measures would no longer be adequate, and the patient would be treated appropriately to adequately address the cardiac decline. The few patients with such aberrant responses properly remain part of the formal treatment group to which they were initially assigned, but they present various special issues, in part because they may not have received a full course of the intended treatment, and in part because they may have experienced some of the worst outcomes observed in the trial. The most basic question is whether one of the treatments causes or prevents such aberrant responses, quite apart from whether that treatment has effects, good or bad, on other patients. Perhaps the aberrant responses were inevitable and would have occurred under either treatment; if so, then the treatments would be judged based on the responses of all patients, with no special focus on aberrant responses. Alternatively, perhaps a typically somewhat

beneficial treatment also causes a few very harmful aberrant effects, and an adequate description of the treatment's effects must mention and distinguish these two aspects, which work in opposite directions. In yet another alternative, perhaps a mildly but typically beneficial treatment confers the additional benefit that it prevents a few very harmful aberrant effects, and possibly the second benefit, rare though it may be, is of comparable clinical significance to the smaller typical benefit.

Common analyses of randomized experiments proceed along one of two paths. First, aberrant responses simply may be included along with more typical responses, perhaps with the aberrant responses as the largest ranks in Wilcoxon's rank-sum test. This approach may be useful as a first step, and it is technically correct as a test of the null hypothesis of no effect of any kind, but it does not distinguish clinically relevant typical effects from perhaps life-threatening aberrant effects. In the second approach, the presence or absence of an aberrant response is noted, and this binary response is examined using Fisher's exact test for a 2×2 table. This isolates aberrant effects and is technically correct as a test of the hypothesis that the treatments do not differ in causing aberrant effects, but it ignores the magnitude of the aberration. Aberrant responses are typically few in number and heterogeneous in magnitude, so the reduction to a binary response discards quite a bit of useful information in a situation in which useful information is scarce.

Here we propose and illustrate a simple exact nonparametric analysis that isolates aberrant responses but takes the magnitude of the aberration into account. Importantly, despite considering magnitudes, the procedure makes no assumptions about effects that are not aberrant; whatever such effects may be, they do not affect the conclusions. In the illustration, binary analysis by Fisher's exact test provides a marginal result, too weak to inspire much conviction but nonetheless too lopsided to provide any reassurance. The new procedure provides a much sharper conclusion.

The null hypothesis of "no aberrant effects" asserts that the aberrant responses were inevitable and would have occurred in

Paul R. Rosenbaum is Professor, Department of Statistics (E-mail: rosenbaum@stat.wharton.upenn.edu), and Jeffrey H. Silber is Professor, Department of Pediatrics, University of Pennsylvania, Philadelphia, PA 19104. Rosenbaum was supported by grant SES-0646002 from the National Science Foundation and Silber was supported by grant R01 HL-50424 from the National Heart, Lung and Blood Institute and grant CA-16520 from the National Cancer Institute.

the same way under either treatment. The alternative is that the treatments differ in causing aberrant responses. The null hypothesis is an instance of a dividing hypothesis, in the sense discussed by Cox (1977, sec. 2.5); it divides alternatives in which one treatment causes an excess of aberrant responses from alternatives in which the other treatment causes the excess. Rejection of a such dividing hypothesis using an appropriate two-sided test provides evidence of the direction of the effect and constitutes rejection also of a set of alternative hypotheses pointing in the opposite direction (see Cox 1977, secs. 2.5 and 4.2). Failure to reject such a dividing hypothesis cautions investigators that they are unable to determine even the direction of the effect, if any.

1.2 Example: The ACE Inhibitor After Anthracycline Randomized Trial

Although anthracyclines are quite effective at curing cancers of childhood, perhaps half of survivors show cardiac abnormalities 10–20 years after cancer diagnosis. With the aim of preventing cardiac decline after treatment with anthracyclines, the AAA randomized trial examined the possible benefits of a second drug, enalapril (Silber et al. 2001, 2004; Greevy et al. 2004). There were $I = 135$ children at least 8 years old who had developed cancer before the age of 20 and who had experienced cardiac decline at least 4 years after cancer diagnosis and at least 2 years after completing cancer treatment. Of these, $n = 69$ were randomly assigned to enalapril, and the remaining $I - n = 66$ received a double-blind placebo. Patients entered the trial gradually and received cardiac performance testing every 6 months, with more than 7 years of follow-up for some of the first patients entered.

As is standardly done in clinical trials of human subjects, the trial's design protocol listed specific criteria for removing a patient from treatment if the patient's condition deteriorated to the point that special treatment was required. For instance, if a patient developed certain acute forms of congestive heart failure, or if the echocardiographic left ventricular shortening fraction (SF) declined by more than 20%, or if the maximum cardiac index on exercise testing (MCI) declined by more than 30%, then the patient would be removed from the study and treated appropriately (see Silber et al. 2004, p. 822, for a complete, detailed specification of these criteria and calculations). A matter of some concern in interpreting the trial's results was that of the seven patients whose protocol treatment was discontinued because of substantial cardiac decline, six were in the placebo group, and one of these six subsequently died of congestive heart failure (see Silber et al. 2004, table 5). The frequencies of discontinuation due to deterioration were 1/69 in the enalapril group and 6/66 in the placebo group. Comparing these using Fisher's exact test for a 2×2 table, the one-sided significance level is .051, and, for a two-sided test, $2 \times .051 = .102$. A similar analysis was done by Silber et al. (2004, p. 825). Of course, this analysis makes no use of information about the magnitude of the deterioration.

For six of these seven patients (including the patient on enalapril), the MCI was higher at the end of the study treatment than at baseline, suggesting improvement, whereas the SF declined in six of seven patients, suggesting deterioration. The SFs are given in Table 1. Patient 1 received enalapril and had the

Table 1. Left ventricular shortening fraction for 7 of 135 patients removed from study treatment due to cardiac decline

Patient	1	2	3	4	5	6	7
Treatment	E	P	P	P	P	P	P
Baseline SF	30.0	27.8	30.3	34.9	35.0	23.2	27.0
End of study treatment SF	25.5	22.2	23.2	26.5	28.0	17.8	29.1
Decline in SF	4.5	5.6	7.1	8.4	7.0	5.4	-2.1
% Decline in SF	15.1	20.1	23.4	24.1	20.0	23.3	-7.8

NOTE: E, enalapril; P, placebo. Source: Silber et al. (2004, p. 825).

second smallest decline in SF, behind only patient 7. Patient 7 experienced the only increase in SF but also was the patient who subsequently died of congestive heart failure. The proposed test takes into account both the number and magnitudes of aberrant responses.

1.3 Outline: Tests, Deciding What to Test, and Confidence Intervals

The article is organized as follows. Section 2 defines and tests the null hypothesis of no aberrant effect of treatment. The null hypothesis of no aberrant effect may be true when the null hypothesis of no effect is false. In Section 1.2 we focused on one aspect, the shortening fraction or SF, which was responsible for five of the seven aberrant responses; in Section 3 we ask whether this raises issues of implicitly testing multiple hypotheses. Is it necessary to correct for multiple testing, say using the Bonferroni inequality? (It is not.) Confidence intervals for the magnitude of the aberrant effect are proposed in Section 4, but, unfortunately, these are applicable only when the study protocol defines an aberrant response in a very simple way. Finally, a brief summary and discussion are provided in Section 5, where we suggest that the protocol for future clinical trials should contain an explicit plan for the analysis of aberrant effects.

2. DO TREATMENTS CAUSE ABERRANT EFFECTS?

2.1 Review: Randomized Experiments and Effects Caused by Treatments

In a randomized experiment, there are I patients, $i = 1, \dots, I$, of whom n are picked at random to receive the treatment, $1 \leq n < I$, whereas the remaining $I - n$ patients receive the control. In Section 1.2, there were $I = 135$ patients; $n = 69$ received enalapril, and $I - n = 66$ received placebo. Write $Z_i = 1$ if i receives treatment, $Z_i = 0$ if i receives control, and $\mathbf{Z} = (Z_1, \dots, Z_I)^T$.

Each patient has two potential p -dimensional vector responses, the response \mathbf{r}_{Ti} that i would exhibit under treatment and the response \mathbf{r}_{Ci} that i would exhibit under control (see Neyman 1923; Rubin 1974). The observed response \mathbf{R}_i from i is $\mathbf{R}_i = \mathbf{r}_{Ti}$ if i receives treatment, $Z_i = 1$, or $\mathbf{R}_i = \mathbf{r}_{Ci}$ if i receives control, $Z_i = 0$, so that $\mathbf{R}_i = Z_i \mathbf{r}_{Ti} + (1 - Z_i) \mathbf{r}_{Ci}$. For patient i , the effect caused by the treatment is a comparison of what would have happened under treatment, \mathbf{r}_{Ti} , and what would have happened under control, \mathbf{r}_{Ci} , such as $\mathbf{r}_{Ti} - \mathbf{r}_{Ci}$. But because \mathbf{r}_{Ti} and \mathbf{r}_{Ci} are never both observed for the same patient i , causal effects cannot be calculated from observed data. Write $\mathcal{F} = \{(\mathbf{r}_{Ti}, \mathbf{r}_{Ci}), i = 1, \dots, I\}$ for the potential responses

of the I patients under treatment and control; thus causal effects $\mathbf{r}_{Ti} - \mathbf{r}_{Ci}$ are functions of \mathcal{F} . The null hypothesis of no treatment effect of any kind asserts that

$$H_0^* : \mathbf{r}_{Ti} = \mathbf{r}_{Ci}, \quad i = 1, \dots, I; \quad (1)$$

that is, patients have varied responses, but each patient i would exhibit the same response under treatment and under control. Here the hypothesis H_0^* is an assertion about \mathcal{F} .

In Section 1.2, $(\mathbf{r}_{Ti}, \mathbf{r}_{Ci})$ contains longitudinal data for patient i 's cardiac condition, including MCI, SF, wall stress, stress velocity index, and the occurrence of congestive heart failure, under treatment and under control. (See Greevy et al. 2004 for a longitudinal analysis of typical effects on wall stress using this notation.)

There are $\binom{I}{n}$ possible values \mathbf{z} of the treatment assignment \mathbf{Z} ; place them in a set Ω , so that $\mathbf{z} \in \Omega$ if and only if $\mathbf{z} = (z_1, \dots, z_I)^T$ with each z_i equal to 0 or 1 and $n = \sum_{i=1}^I z_i$. Write $|S|$ for the number of elements of a finite set S , so that $|\Omega| = \binom{I}{n}$. In a randomized experiment, one assignment \mathbf{Z} is picked at random from Ω using random numbers, so that $\Pr(\mathbf{Z} = \mathbf{z} | \mathcal{F}) = \binom{I}{n}^{-1} = 1/|\Omega|$ for each $\mathbf{z} \in \Omega$.

In Fisher's (1935) theory of randomization inference, probability enters only through the random assignment of treatments, \mathbf{Z} , whose distribution is known because it was created by the experimenter using random numbers. In this way, randomization creates the needed probability distributions and forms "the reasoned basis for inference" in Fisher's phrase. More precisely, even though \mathcal{F} is neither known nor observable, randomization ensures that $\Pr(\mathbf{Z} = \mathbf{z} | \mathcal{F}) = \binom{I}{n}^{-1}$ for each $\mathbf{z} \in \Omega$. If \mathbf{e} is any function of \mathcal{F} and $t(\mathbf{Z}, \mathbf{e})$ is any statistic, then the randomization distribution of the statistic, $\Pr\{t(\mathbf{Z}, \mathbf{e}) \geq v | \mathcal{F}\}$, is simply the proportion of treatment assignments $\mathbf{z} \in \Omega$ yielding $t(\mathbf{z}, \mathbf{e}) \geq v$, that is, $\Pr\{t(\mathbf{Z}, \mathbf{e}) \geq v | \mathcal{F}\} = |\{\mathbf{z} \in \Omega : t(\mathbf{z}, \mathbf{e}) \geq v\}| / |\Omega|$. Because \mathcal{F} is unobservable, for most definitions of \mathbf{e} , neither $t(\mathbf{Z}, \mathbf{e})$ nor $\Pr\{t(\mathbf{Z}, \mathbf{e}) \geq v | \mathcal{F}\}$ can be determined; however, the situation may be different when \mathcal{F} satisfies certain null hypotheses. For instance, if H_0^* in (1) were true, then $\mathbf{r}_{Ti} = \mathbf{r}_{Ci} = \mathbf{R}_i$ is always observed, so that both $t(\mathbf{Z}, \mathbf{e})$ and $\Pr\{t(\mathbf{Z}, \mathbf{e}) \geq v | \mathcal{F}\}$ may be calculated for any function \mathbf{e} of \mathcal{F} (see Rosenbaum 2005, sec. 2.1, and Small, Ten Have, and Rosenbaum 2008 for elaboration this description of Fisher's randomization tests). For example, if the response is one-dimensional and binary, then Fisher's (1935) exact test for a 2×2 table uses the randomization distribution to test (1), whereas if the response is one-dimensional and continuous, then the exact distribution of Wilcoxon's rank-sum statistic is a randomization distribution used to test (1) (see, e.g., Lehmann 1998, sec. 1).

2.2 Hypotheses That Focus on Aberrant Effects, Apart From Other Effects

There is a subset, \mathcal{A} , of p -dimensional space that defines an aberrant or abnormal response, and $(\mathbf{r}_{Ti}, \mathbf{r}_{Ci})$ contains the potential responses used in the definition of an aberrant response. In Section 1.2, \mathcal{A} is defined in terms of certain magnitudes of change from baseline in MCI or SF or the occurrence of congestive heart failure, but Silber et al. (2004, p. 821) provided the exact definition of \mathcal{A} , which makes use of confirming, repeated measurements. In this notation, i exhibits an aberrant response if $\mathbf{R}_i \in \mathcal{A}$. Exposure to treatment would cause i to have

an aberrant response that would not be aberrant under control if $\mathbf{r}_{Ti} \in \mathcal{A}$ but $\mathbf{r}_{Ci} \notin \mathcal{A}$. Exposure to treatment did cause i to have an aberrant response if $\mathbf{r}_{Ti} \in \mathcal{A}$, $\mathbf{r}_{Ci} \notin \mathcal{A}$, and $Z_i = 1$, so that $\mathbf{R}_i \in \mathcal{A}$. It may happen that i would have an aberrant response whether treated or not, but the response would be different, $\mathbf{r}_{Ti} \in \mathcal{A}$, $\mathbf{r}_{Ci} \in \mathcal{A}$, $\mathbf{r}_{Ti} \neq \mathbf{r}_{Ci}$. (More generally, \mathcal{A} could be defined in terms of both posttreatment responses, \mathbf{R}_i , and observed pretreatment covariates, say \mathbf{x}_i ; however, this introduces no new technical issues and makes the notation more complex, so this possibility will not be explicit in the notation.)

A null hypothesis of interest is that the treatment and control do not differ in their effects for patients with aberrant responses, so that, in particular, a patient's serious deterioration would have occurred in the same way under treatment or under control. Formally, the null hypothesis of no aberrant effect states that

$$H_0 : \mathbf{r}_{Ti} = \mathbf{r}_{Ci} \text{ if either } \mathbf{r}_{Ti} \in \mathcal{A} \text{ or } \mathbf{r}_{Ci} \in \mathcal{A}, \quad i = 1, \dots, I, \quad (2)$$

which is again an assertion about \mathcal{F} . The null hypothesis of no treatment effect of any kind (1) implies the null hypothesis of no aberrant effect (2), but the converse is untrue; thus (2) is a weaker hypothesis in that it asserts much less about \mathcal{F} . In particular, (2) says nothing about the effects of the treatment on patients i who would have aberrant responses under neither treatment nor control, $\mathbf{r}_{Ti} \notin \mathcal{A}$ and $\mathbf{r}_{Ci} \notin \mathcal{A}$; that is, (2) says nothing about the effects on most patients. We test (2) to appraise evidence about three possibilities: (a) Treatment produces more extensive aberrant responses than control, (b) control produces more extensive aberrant responses than treatment, or (c) the evidence available is insufficient to distinguish (a) and (b).

If (2) is false, then treatment and control have different effects on aberrant responses. This can occur for each patient i in three ways. If $\mathbf{r}_{Ti} \in \mathcal{A}$, $\mathbf{r}_{Ci} \notin \mathcal{A}$, then $\mathbf{r}_{Ti} \neq \mathbf{r}_{Ci}$ and i would have an aberrant response under treatment but not under control. If $\mathbf{r}_{Ti} \notin \mathcal{A}$, $\mathbf{r}_{Ci} \in \mathcal{A}$, then i would have an aberrant response under control but not under treatment. Finally, if $\mathbf{r}_{Ti} \in \mathcal{A}$ and $\mathbf{r}_{Ci} \in \mathcal{A}$ but $\mathbf{r}_{Ti} \neq \mathbf{r}_{Ci}$, then i would have an aberrant response under treatment and under control, but the effect $\mathbf{r}_{Ti} - \mathbf{r}_{Ci}$ on this patient would be different.

Consider, for instance, the use of Fisher's exact test applied to (2). Write $y_{Ti} = 1$ if $\mathbf{r}_{Ti} \in \mathcal{A}$, $y_{Ti} = 0$ if $\mathbf{r}_{Ti} \notin \mathcal{A}$, and $y_{Ci} = 1$ if $\mathbf{r}_{Ci} \in \mathcal{A}$, $y_{Ci} = 0$ if $\mathbf{r}_{Ci} \notin \mathcal{A}$. If the null hypothesis of no aberrant effect (2) is true, then the (y_{Ti}, y_{Ci}) 's satisfy the null hypothesis of no effect of any kind, that is, $H_0 : y_{Ti} = y_{Ci}$, $i = 1, \dots, I$, which is, of course, the hypothesis tested by Fisher's exact test. Although converting \mathbf{R}_i to a binary indicator in this way discards information about magnitudes, it does succeed in testing the weaker null hypothesis of no aberrant effect (2) rather than the hypothesis of no effect at all (1). Can an exact randomization test of (2) rather than (1) be constructed that uses more information?

To generalize this reasoning, now let y_{Ti} be an aspect, not necessarily binary, of \mathbf{r}_{Ti} , and y_{Ci} be an aspect of \mathbf{r}_{Ci} , with observed value $Y_i = y_{Ti}$ if $Z_i = 1$ and $Y_i = y_{Ci}$ if $Z_i = 0$. An "aspect" may be any single coordinate of the response or any other function of the response, possibly a vector-valued function; formally, there is a function $f(\cdot)$ with $y_{Ti} = f(\mathbf{r}_{Ti})$,

$y_{Ci} = f(\mathbf{r}_{Ci})$, $i = 1, \dots, I$, so (y_{Ti}, y_{Ci}) is a function of \mathcal{F} . Consider the null hypothesis

$$H_0^y : y_{Ti} = y_{Ci} \text{ if either } \mathbf{r}_{Ti} \in \mathcal{A} \text{ or } \mathbf{r}_{Ci} \in \mathcal{A}, \quad i = 1, \dots, I, \quad (3)$$

which asserts that there is no effect on the aspect (y_{Ti}, y_{Ci}) for patients who would have an aberrant response under treatment or control or both. If the aspect is the entire response, $f(\mathbf{r}_{Ti}) = \mathbf{r}_{Ti}$ and $f(\mathbf{r}_{Ci}) = \mathbf{r}_{Ci}$, then (2) and (3) are the same. However, in general, H_0 in (2) implies H_0^y in (3), but not conversely. With different definitions $f(\cdot)$ of the aspect (y_{Ti}, y_{Ci}) , a variety of hypotheses may be expressed in the form H_0^y in (3). Write $\mathbf{Y} = (Y_1, \dots, Y_I)^T$.

Suppose that M patients are observed to have aberrant responses, $\mathbf{R}_i \in \mathcal{A}$. In Table 1, Y_i is the change in SF from baseline, and $M = 7$. Rank the Y_i in the following way. For the M patients with aberrant responses, rank their M observed Y_i 's from 1 to M with average ranks for ties. For the remaining $I - M$ patients, assign rank 0. Let Q_i , $i = 1, \dots, I$, be the resulting ranks for the I patients, and write $\mathbf{Q} = (Q_1, \dots, Q_I)^T$. When sorted into increasing order, if there are no ties, then the I ranks are $(0, 0, \dots, 0, 1, 2, \dots, M)$, where there are $b = I - M$ 0's. The test statistic is the sum of the ranks in the treated group, $A = \mathbf{Z}^T \mathbf{Q}$, which can range from 0 to $M(M + 1)/2 = 1 + 2 + \dots + M$, or from 0 to 28 in Section 1.2. A closely related test statistic was proposed by Mehrotra, Li, and Gilbert (2006) for tests about the "burden of illness" concept of Chang, Guess, and Heysse (1994).

Proposition 1. Suppose that H_0^y in (3) is true. Then \mathbf{Q} is fixed given \mathcal{F} , and the randomization distribution $\Pr(A \geq a | \mathcal{F})$ of $A = \sum_{i=1}^I Z_i Q_i$ is that of the sum of n numbers selected at random without replacement from the I coordinates of $\mathbf{Q} = (Q_1, \dots, Q_I)^T$.

Proof. Divide the I patients into two disjoint subsets, $\{1, 2, \dots, I\} = \mathcal{S}_0 \cup \mathcal{S}_1$, $\emptyset = \mathcal{S}_0 \cap \mathcal{S}_1$, where $\mathcal{S}_0 = \{i : \mathbf{r}_{Ti} \notin \mathcal{A} \text{ and } \mathbf{r}_{Ci} \notin \mathcal{A}\}$ and $\mathcal{S}_1 = \{i : \mathbf{r}_{Ti} \in \mathcal{A} \text{ or } \mathbf{r}_{Ci} \in \mathcal{A}\}$; this division is a function of \mathcal{F} . If $i \in \mathcal{S}_0$, then $\mathbf{R}_i = Z_i \mathbf{r}_{Ti} + (1 - Z_i) \mathbf{r}_{Ci} \notin \mathcal{A}$, so $Q_i = 0$. Using (3), if $i \in \mathcal{S}_1$, then $y_{Ti} = y_{Ci} = Y_i$; thus for $i \in \mathcal{S}_1$, the value of Y_i is determined by \mathcal{F} . Therefore, the rank Q_i of Y_i among the Y_j , $j \in \mathcal{S}_1$, is also determined by \mathcal{F} . It follows that the randomization distribution, $\Pr(\mathbf{Z} = \mathbf{z} | \mathcal{F}) = \binom{I}{n}^{-1}$, picks n scores at random without replacement from the I scores $\mathbf{Q} = (Q_1, \dots, Q_I)^T$, which are fixed conditionally given \mathcal{F} .

Remark 1. In the definition of \mathbf{Q} and $A = \mathbf{Z}^T \mathbf{Q}$, it is important that patients without aberrant responses receive rank 0, $Q_i = 0$ if $\mathbf{R}_i \notin \mathcal{A}$, but any nonnegative ranks $0 \leq d_1 \leq d_2 \leq \dots \leq d_M$ may be used to rank the M aberrant responses; the proof of Proposition 1 is unchanged. In particular, for patient i , if $\mathbf{R}_i \in \mathcal{A}$ but the aspect Y_i is not the reason for this, as is true of SF for patients 1 and 7 in Table 1, then there is the option of defining the corresponding rank Q_i to be 0. We did not do this, because we prefer to view the 15.1% decline in SF for patient 1 in Table 1 as worse than the 7.8% improvement in SF for patient 7. Also, for a multivariate Y_i , one of several standard multivariate ranking methods may be used for the M patients with aberrant responses; see Section 4.4 for specifics.

For the decline in SF in Table 1, $A = 2$ because there are $M = 7$ aberrant responses, one of which was from a treated patient, $Z_i = 1$, and that patient's change in SF was the second to smallest, $Q_i = 2$. When the number of aberrant responses, M , is small, as it typically is even in large clinical trials, the exact null distribution is easy to obtain; see Section 2.3 for specifics.

In particular cases, A becomes either Wilcoxon's rank-sum test or Fisher's exact test for a 2×2 table. If $\mathcal{A} = \mathbb{R}^p$, so that all responses are always aberrant, then (1) and (2) are equivalent; thus, without ties, the sorted ranks are $(1, 2, \dots, I)$, and A is Wilcoxon's rank-sum statistic. If (y_{Ti}, y_{Ci}) is the binary indicator of an aberrant response, $y_{\ell i} = 1$ if $\mathbf{r}_{\ell i} \in \mathcal{A}$, and $y_{\ell i} = 0$ if $\mathbf{r}_{\ell i} \notin \mathcal{A}$, for $\ell = T, C$, then all M of the Y_i for aberrant responses are equal, so the sorted average ranks are $(0, \dots, 0, (M + 1)/2, \dots, (M + 1)/2)$, and A produces the same significance levels as Fisher's exact test for a 2×2 table.

The statistic A has a simple form, but the randomization distribution of any test of (3) is determined in a similar way. Let $Y_i^* = Y_i$ if $\mathbf{R}_i \in \mathcal{A}$ and $Y_i^* = *$ otherwise, so $Y_i^* \in \mathbb{R} \cup \{*\} = \mathbb{R}^*$, say. Let $\mathbf{Y}^* = (Y_1^*, \dots, Y_I^*)^T$, so $\mathbf{Y}^* \in \mathbb{R}^* \times \dots \times \mathbb{R}^* = \mathbb{R}^{*I}$, say. In words, \mathbf{Y}^* records the value of the aspect Y_i for patients with aberrant responses and records $*$ for other patients. Let $t(\mathbf{Y}^*, \mathbf{Z})$ be any test statistic, that is, any function $t : \mathbb{R}^{*I} \times \Omega \rightarrow \mathbb{R}$. Of course, A is one such statistic. Proposition 2 generalizes Proposition 1, but the proof is similar. In considering Proposition 2, it should be emphasized that if (3) is true, then \mathbf{Y}^* is a function of \mathcal{F} and thus is fixed by conditioning on \mathcal{F} , but \mathbf{Y} remains a function of \mathcal{F} and \mathbf{Z} jointly, and so \mathbf{Y} remains a nondegenerate random variable given \mathcal{F} .

Proposition 2. Suppose that H_0^y in (3) is true. Then \mathbf{Y}^* is a function of \mathcal{F} , so the randomization distribution of $t(\mathbf{Y}^*, \mathbf{Z})$ is

$$\Pr(t(\mathbf{Y}^*, \mathbf{Z}) \geq a | \mathcal{F}) = \frac{|\mathbf{z} \in \Omega : t(\mathbf{Y}^*, \mathbf{z}) \geq a|}{|\Omega|}.$$

2.3 Exact Null Distribution

Suppose that the I ranks are $\mathbf{Q}^{(I)} = (0, 0, \dots, 0, d_1, \dots, d_M)^T$, with $b \geq 0$ 0's followed by $M \geq 0$ positive, nondecreasing ranks, $0 < d_1 \leq d_2 \leq \dots \leq d_M$, with $b + M = I$. Write $h_I\{a, n, \mathbf{Q}^{(I)}\}$ for the number of treatment assignments $\mathbf{z} \in \Omega$ such that $\mathbf{z}^T \mathbf{Q}^{(I)} = a$. There is a simple recursion formula that writes $h_I\{a, n, \mathbf{Q}^{(I)}\}$ in terms of a smaller problem with $I - 1$ patients, having $I - 1$ ranks $\mathbf{Q}^{(I-1)} = (0, 0, \dots, 0, d_1, \dots, d_{M-1})$, again with b 0's. Two possibilities yield $\mathbf{z}^T \mathbf{Q}^{(I)} = a$: If $z_I = 0$, then $a = \sum_{i=1}^{I-1} z_i Q_i$, or if $z_I = 1$, then $a - Q_I = \sum_{i=1}^{I-1} z_i Q_i$. This yields

$$h_I\{a, n, \mathbf{Q}^{(I)}\} = h_{I-1}\{a, n, \mathbf{Q}^{(I-1)}\} + h_{I-1}\{a - Q_I, n - 1, \mathbf{Q}^{(I-1)}\}, \quad (4)$$

with initial conditions

$$\begin{aligned} h_I\{a, n, \mathbf{Q}^{(I)}\} &= 0 \quad \text{if } a < 0 \text{ or } n < 0 \text{ or } I < n, \\ h_I\{0, n, \mathbf{Q}^{(I)}\} &= \binom{b}{n} \quad \text{if } I = b, \quad \text{and} \\ h_I\{a, n, \mathbf{Q}^{(I)}\} &= 0 \quad \text{if } a \neq 0 \text{ and } I = b. \end{aligned} \quad (5)$$

If $\mathbf{Q} = (0, 0, \dots, 0, 1, \dots, 1)^T$, then the recursion (4) generates significance levels for Fisher's exact test for a 2×2 table,

whereas if $b = 0$ and $\mathbf{Q} = (1, 2, \dots, I)^T$, it generates the distribution of Wilcoxon's rank-sum test. Because of shortcut (5), which eliminates many steps when b is large, the recursion (4) runs very quickly for large I providing that $M = I - b$ is small, for instance, $M = 7$ in the example. R code implementing (4) is given in the Appendix. Without the shortcut (5), related recursions are well known in the literature on rank tests (e.g., Brunner 1991, p. 1150).

2.4 Example: Aberrant Effects in the AAA Trial

For the example given in Section 1.2, Y_i is SF, the statistic A assigns ranks 0 to the $b = I - M = 135 - 7 = 128$ patients who were not removed from the trial due to deterioration, and ranks 1 to 7 to the $M = 7$ declines in SF for the 7 removed patients. Then the statistic is actually $A = 2$, but A could have taken any integer value between 0 and $28 = 1 + 2 + \dots + 7$.

Table 2 is produced with straightforward, quick calculations. There are $M = 7$ aberrant responses among $I = 135$ patients, of whom $n = 69$ were in the treated group; thus the sorted ranks are $(0, \dots, 0, 1, 2, 3, 4, 5, 6, 7)$. Of the $\binom{135}{69}$ treatment assignments $\mathbf{z} \in \Omega$, there are $\binom{128}{69}$ that yield $A = 0$, $\binom{128}{68}$ that yield $A = 1$, and $\binom{128}{68}$ that yield $A = 2$; thus the one-sided significance level is

$$\Pr(A \leq 2) = \frac{\binom{128}{69} + \binom{128}{68} + \binom{128}{68}}{\binom{135}{69}} = .0186.$$

In comparison, for Fisher's exact test, there are $\binom{128}{69}$ treatment assignments that place all 7 aberrant responses in the placebo group, and $\binom{7}{1} \binom{128}{68} = 7 \times \binom{128}{68}$ treatment assignments with 1 aberrant response in the enalapril group and 6 in the placebo group, yielding a one-sided significance level of

$$\frac{\binom{128}{69} + 7 \times \binom{128}{68}}{\binom{135}{69}} = .0509.$$

The distribution in Table 2 is discrete and asymmetric because $n \neq I - n$, and in such a situation there are various notions of a two-sided significance level, the simplest being $2 \times .0186 = .0372$. There are, however, arguments for defining the two-sided significance level as the "the one-sided level q_{obs} plus the one-sided level from the other tail nearest to but not exceeding q_{obs} " (see Cox 1977, sec. 4.2; Cox and Hinkely 1974, p. 79). In Table 2, this two-sided significance level for $A = 2$ is then $.0186 + .0160 = .0346$. With either definition, the two-sided significance level from A is smaller than the one-sided significance level from Fisher's exact test.

Table 2. Tails of the null randomization distribution of A with $I = 135$ patients, $n = 69$ on treatment, and $M = 7$ aberrant responses

a	$\Pr(A \leq a)$	a	$\Pr(A \geq a)$
0	.0056	28	.0078
1	.0121	27	.0160
2	.0186	26	.0241
3	.0322	25	.0406
4	.0459	24	.0570
5	.0668	23	.0818
6	.0955	22	.1147

3. MULTIPLE ASPECTS: SELECTING THE HYPOTHESIS TO TEST

In Section 1.2 the set \mathcal{A} that defined an aberrant response used three outcomes—declines in SF, MCI and specific forms of congestive heart failure—with $M = 7$ patients exhibiting an aberrant response, but the test in Section 2.4 focused on SF alone. Recall from Section 1.2 that SF deteriorated for six of seven patients, whereas MCI improved for six of seven patients. Is there an implicit problem of multiple testing here? Can one decide to focus on one aspect, say SF, rather than another aspect, say MCI, after noting that patients' responses most often tended to be aberrant because of SF rather than because of MCI? Or does such a decision invalidate the test in Section 2.4? Once stated formally, this question has a simple answer.

Suppose that there are K scalar aspects defined in the protocol before the experiment begins, $y_{Tik} = f_k(\mathbf{r}_{Ti})$, $y_{Cik} = f_k(\mathbf{r}_{Ci})$, and $Y_{ik} = Z_i y_{Tik} + (1 - Z_i) y_{Cik}$, $k = 1, \dots, K$, with $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iK})^T$, \mathbf{y}_{Ti} , and \mathbf{y}_{Ci} the corresponding K -dimensional vectors. As an illustration in the example of Section 1.2, let $K = 2$, with the two aspects being the declines in SF for $k = 1$ and MCI for $k = 2$, with decline defined as baseline minus end-of-study treatment.

For each aspect, there is a subset of the real line, \mathcal{Y}_k , defined in the protocol, of relevant values of Y_{ik} , so patient i has a relevant value of aspect k if $Y_{ik} \in \mathcal{Y}_k$. For instance, with $\mathcal{Y}_1 = \mathcal{Y}_2 = \{y : y > 0\}$, any decline in SF or MCI is relevant. In Section 1.2 only substantial declines in SF or MCI constituted an aberrant response, but here any decline is counted as relevant; the terms "aberrant" and "relevant" need not coincide. In Section 1.2, 6 of 7 patients with aberrant responses, $\mathbf{R}_i \in \mathcal{A}$, had relevant declines in SF, $Y_{i1} \in \mathcal{Y}_1$, but only 1 of 7 had a relevant decline in MCI, $Y_{i2} \in \mathcal{Y}_2$. Define

$$W_k = \{|i : \mathbf{R}_i \in \mathcal{A}, Y_{ik} \in \mathcal{Y}_k\},$$

so that W_k is the number of patients with aberrant responses having relevant values of aspect k . In Section 1.2, $W_1 = 6$ and $W_2 = 1$, and we decided to focus on SF rather than MCI because $W_1 > W_2$. It is important to note that W_k counts relevant responses, $Y_{ik} \in \mathcal{Y}_k$, but ignores whether they occur in the treated or control group; this is essential.

Let $w(\cdot)$ be a rule or function defined in the protocol that looks at W_1, \dots, W_K to select one aspect, $L \in \{1, 2, \dots, K\}$, as the focus of attention, that is, $w(W_1, \dots, W_K) = L$. The test in Section 2.2 is then applied just to aspect Y_{iL} . One simple rule is to pick k so that $W_k = \max_{1 \leq j \leq K} W_j$, and if there is a tie, if several k achieve the maximum, then pick the smallest such k , because the protocol ordered the aspects by an a priori sense of relative importance. In general, L is a random variable, because it is calculated from \mathbf{R}_i and Y_{ik} , which depend on both \mathbf{Z} and \mathcal{F} .

If the null hypothesis (2) were true, then Proposition 1 says that the ranks \mathbf{Q}_k for any one aspect k , with k chosen a priori, are functions of \mathcal{F} , not varying with $\mathbf{Z} \in \Omega$, so that the null randomization distribution $\Pr(A_k \geq a | \mathcal{F})$ of $A_k = \mathbf{Z}^T \mathbf{Q}_k$ given \mathcal{F} is simply the distribution of the sum of n ranks picked by simple random sampling without replacement from the I fixed coordinates of \mathbf{Q}_k . What about the distribution of A_L where $L = w(W_1, \dots, W_K)$ is picked after examining the data in the precise sense defined earlier? In general, A_L is picked from the

random variables A_1, \dots, A_K based on the random variable L ; thus in general the distribution of A_L is not the distribution of any one of the A_k 's. Proposition 3 says situation is simpler under the null hypothesis of no aberrant effect (2): then the choice of $L = k$ is fixed by \mathcal{F} and would be the same choice for every random assignment of treatments $\mathbf{Z} \in \Omega$.

Proposition 3. If the null hypothesis of no aberrant effect (2) is true, then $L = w(W_1, \dots, W_K)$ is a function of \mathcal{F} , taking the same value for all $\mathbf{Z} \in \Omega$, and

$$\Pr(A_L \geq a | \mathcal{F}) = \Pr(A_L \geq a | \mathcal{F}, L = k) = \Pr(A_k \geq a | \mathcal{F})$$

if the realized value of L is k .

Proof. Assume that (2) is true. Let $\mathcal{I} = \{i: \mathbf{R}_i \in \mathcal{A}\}$ be the set of patients with aberrant responses. If $\mathbf{r}_{Ti} \notin \mathcal{A}$ and $\mathbf{r}_{Ci} \notin \mathcal{A}$, then $\mathbf{R}_i \notin \mathcal{A}$ and $i \notin \mathcal{I}$. If either $\mathbf{r}_{Ti} \in \mathcal{A}$ or $\mathbf{r}_{Ci} \in \mathcal{A}$, then, by (2), it follows that $\mathbf{r}_{Ti} = \mathbf{r}_{Ci} = \mathbf{R}_i$, $i \in \mathcal{I}$, and $y_{Tik} = y_{Cik} = Y_{ik}$, for $k = 1, \dots, K$. Therefore, (2) implies that

$$\begin{aligned} \{i: \mathbf{R}_i \in \mathcal{A}, Y_{ik} \in \mathcal{Y}_k\} &= \{i: \mathbf{r}_{Ti} \in \mathcal{A}, y_{Tik} \in \mathcal{Y}_k\} \\ &= \{i: \mathbf{r}_{Ci} \in \mathcal{A}, y_{Cik} \in \mathcal{Y}_k\}, \end{aligned}$$

so this set is determined by \mathcal{F} , not changing with $\mathbf{Z} \in \Omega$, so each W_k and thus also $L = w(W_1, \dots, W_K)$ are determined by \mathcal{F} . This proves $\Pr(A_L \geq a | \mathcal{F}) = \Pr(A_L \geq a | \mathcal{F}, L = k) = \Pr(A_k \geq a | \mathcal{F}, L = k) = \Pr(A_k \geq a | \mathcal{F})$, as required.

In short, in Section 2.4 the test using SF gave a correct significance level for testing (2) even though we decided to look at SF rather than MCI because there were more declines in SF. Intuitively, (W_1, W_2) counts declines in SF and MCI, but under the null hypothesis (2), (W_1, W_2) contains no information about whether those declines favor the treatment or the control.

A word should be said about patient 7 in Table 1. Our focus on SF counts this placebo patient's increase (negative decline) in SF as the most favorable result among the seven aberrant responses, even though this patient subsequently died of congestive heart failure. Had we counted this outcome as extremely unfavorable rather than extremely favorable, then in Section 2.4, the statistic would have been $A = 1$ with two-sided significance level $.0121 + .0078 = .0199$, rather than $.0346$ for $A = 2$ in Section 2.4. Because the statistical method we are illustrating is new, the protocol that specified the design and planned analyses for the AAA study did not propose a specific analysis of aberrant effects using this statistic, so it is debatable which of these two analyses is more appropriate. The best approach in a future use of the method would be to include in the protocol a specific plan for ranking aberrant responses according to their severity. That plan could be an adaptive approach of the type suggested by Proposition 3, or it could combine several aspects into a single index and focus on that one index, or it could apply the adaptive approach to K aspects that include both some individual measures, like SF and MCI, and one or more summary indices.

4. CONFIDENCE INTERVALS

4.1 Review: Inverting Randomization Tests to Obtain Confidence Intervals for an Additive Treatment Effect

Before discussing confidence intervals for aberrant effects, we briefly review confidence intervals derived by inverting randomization tests. Randomization tests of no treatment effect are

most often inverted to form confidence intervals for an additive treatment effect, say τ , so that treatment increases each patient's response by τ . (See Lehmann 1963, 1998, sec. 2.6, or Rosenbaum 2002, p. 45, for discussion of randomization inference with additive effects, and see Rosenbaum 2001 for nonadditive effects.) In a randomized experiment, an additive effect shifts the observable distribution of responses in the treated group by τ compared with the responses in the control group. Consider testing the null hypothesis $H_0: y_{Ti} = y_{Ci} + \tau_0$, $i = 1, \dots, I$, with some specified shift τ_0 . If the hypothesis were true, then $Y_i = Z_i y_{Ti} + (1 - Z_i) y_{Ci} = y_{Ci} + Z_i \tau_0$, so that the adjusted responses, $Y_i - Z_i \tau_0 = y_{Ci}$ would satisfy the null hypothesis of no effect. Expressed differently, under this null hypothesis, Y_i is not a function of \mathcal{F} , but $Y_i - Z_i \tau_0 = y_{Ci}$ is a function of \mathcal{F} . The hypothesis $H_0: y_{Ti} = y_{Ci} + \tau_0$ is tested by applying a test of no effect, such as Wilcoxon's rank-sum test, to $Y_i - Z_i \tau_0$. The interval of values of τ_0 not rejected in a two-sided .05-level test forms a 95% confidence interval for τ_0 .

In Section 4.2 this reasoning is generalized to the situation in which the aberrant effect is a shift, say δ , with no assumption about the effect on responses that are not aberrant. A key technical point that must be addressed is that a shift of δ may cause a patient to have an aberrant response under one treatment but not under the other.

4.2 Confidence Intervals for Aberrant Treatment Effects

When the set of aberrant responses, \mathcal{A} , has a sufficiently simple definition, it is possible to invert the test in Section 2 obtain a confidence interval for the magnitude of aberrant effects. Specifically, throughout Sections 4.2 and 4.3, assume without further mention that \mathcal{A} is defined just in terms of a scalar Y ; that is, \mathcal{A} is a set of possible values of Y_i , and i has an aberrant response if and only if $Y_i \in \mathcal{A}$; this assumption is removed in Section 4.4. (As in Sec. 2.2, \mathcal{A} may use Y_i and observed pretreatment covariates \mathbf{x}_i ; it simply cannot use responses other than Y_i .)

Consider testing the hypothesis

$$H_0: y_{Ti} = y_{Ci} + \delta_0 \text{ if either } y_{Ti} \in \mathcal{A} \text{ or } y_{Ci} \in \mathcal{A}, \quad i = 1, \dots, I, \quad (6)$$

which reduces to (3) for $\delta_0 = 0$.

Proposition 4. (a) If (6) is true, then the set

$$\mathcal{I}_{\delta_0} = \{i: Y_i \in \mathcal{A}, Y_i - Z_i \delta_0 \in \mathcal{A}, Y_i + (1 - Z_i) \delta_0 \in \mathcal{A}\}$$

is determined by \mathcal{F} .

b. Define $Q_{i, \delta_0} = 0$ if $i \notin \mathcal{I}_{\delta_0}$. If there are $M_{\delta_0} = |\mathcal{I}_{\delta_0}|$ patients i in \mathcal{I}_{δ_0} , then rank their adjusted responses, $Y_i - \delta_0 Z_i$, from 1 to M_{δ_0} with average ranks for ties, and define Q_{i, δ_0} to be this rank for $i \in \mathcal{I}_{\delta_0}$. Write $\mathbf{Q}_{\delta_0} = (Q_{1, \delta_0}, \dots, Q_{I, \delta_0})^T$. If (6) is true, then \mathbf{Q}_{δ_0} is a function of \mathcal{F} .

c. If (6) is true, then the randomization distribution, $\Pr(A_{\delta_0} \geq a | \mathcal{F})$, of $A_{\delta_0} = \mathbf{Z}^T \mathbf{Q}_{\delta_0}$ is the distribution of the sum of n scores picked at random without replacement from the I fixed scores $\mathbf{Q}_{\delta_0} = (Q_{1, \delta_0}, \dots, Q_{I, \delta_0})^T$.

Remark 2. If the null hypothesis (6) is true, then \mathcal{I}_{δ_0} is the set of patients i who would have aberrant responses both under treatment and under control. With $\delta_0 = 0$, the set \mathcal{I}_0 is simply

the set of patients observed to have aberrant responses, $Y_i \in \mathcal{A}$, that is, the patients with nonzero ranks in the test in Section 2.2. For all $\delta_0, \mathcal{I}_{\delta_0} \subseteq \mathcal{I}_0$, so the number of zero rank scores in \mathbf{Q}_{δ_0} in Proposition 4 is always at least as great as the number of zero rank scores in \mathbf{Q} in Section 2.2.

Proof. Suppose that the hypothesis (6) is true. Consider whether or not $i \in \mathcal{I}_{\delta_0}$. There are two cases.

Case 1: If either $y_{Ti} \in \mathcal{A}$ or $y_{Ci} \in \mathcal{A}$, then $y_{Ti} = y_{Ci} + \delta_0$; hence, $Y_i - Z_i\delta_0 = y_{Ci}$ and $Y_i + (1 - Z_i)\delta_0 = y_{Ti}$. In this case, $i \in \mathcal{I}_{\delta_0}$ if and only if $y_{Ci} \in \mathcal{A}$ and $y_{Ti} \in \mathcal{A}$, and this is determined by \mathcal{F} .

Case 2: If $y_{Ti} \notin \mathcal{A}$ and $y_{Ci} \notin \mathcal{A}$, then $Y_i \notin \mathcal{A}$ for all $\mathbf{Z} \in \Omega$, and so $i \notin \mathcal{I}_{\delta_0}$, and this too is determined by \mathcal{F} .

This proves part a. From part a, it follows that if $i \notin \mathcal{I}_{\delta_0}$, then $Q_{i,\delta_0} = 0$. On the other hand, if $i \in \mathcal{I}_{\delta_0}$ then $Y_i - Z_i\delta_0 = y_{Ci}$, so the rank Q_{i,δ_0} of $Y_i - Z_i\delta_0$ is a function of \mathcal{F} , so \mathbf{Q}_{δ_0} is also a function of \mathcal{F} . Then part c follows immediately.

The set of hypotheses (6) not rejected at level α forms a $1 - \alpha$ confidence set. As is typically true with nonparametric confidence sets, when the data are very limited, the confidence set may not be a finite interval, but could be, for instance, a half-line, $(-\infty, v]$ say.

4.3 An Artificial Example

Because the set of aberrant responses, \mathcal{A} , was defined using several criteria and confirming measurements in the AAA trial in Section 1.2, the method in Section 4.2 is not applicable. To provide a brief numerical illustration, suppose that (a) \mathcal{A} had been defined as a decline in SF of 4 or more, (b) patient 7 in Table 1 had a decline of 4.1, and (c) only the 7 patients in Table 1 had aberrant responses, and they are $i = 1, 2, \dots, 7$. These suppositions would leave the test results, $A = 2$, in Section 2.4 unchanged, but would permit the illustration of the method in Section 4.2.

To quickly illustrate the calculations, we first construct a one-sided 97.5% confidence interval. To test (6) with $\delta_0 = -.2$, one computes $\mathcal{I}_{\delta_0} = \{1, 2, \dots, 6\}$, $M_{\delta_0} = 6$, $A_{\delta_0} = 1$, and the one-sided significance level is .0258, so this hypothesis would just barely not be rejected in a one-sided .025 level test. The quantities M_{δ_0} and A_{δ_0} are step functions of δ_0 . For all $\delta_0 < -.1$, $M_{\delta_0} \leq 6$ and $A_{\delta_0} \geq 1$, and the one-sided significance level is $>.025$, whereas for $\delta_0 \geq -.1$, either $M_{\delta_0} = 7$ and $A_{\delta_0} \leq 2$ or $M_{\delta_0} = 6$ and $A_{\delta_0} = 0$, and the same one-sided test rejects these hypotheses with significance levels $<.025$. Thus the one-sided 97.5% confidence interval is $\delta_0 < -.1$, so the artificial data would be incompatible with no aberrant effect, but compatible with trivially small or quite large benefits from treatment.

In the current context, we adopt one of the several standard definitions of a two-sided confidence interval: a two-sided 95% confidence set is the intersection of two 97.5% confidence sets, with 95% coverage for the intersection assured by the Bonferroni inequality. In the current example, the two-sided 95% confidence interval is $(-\infty, -.1)$, which turns out to be the same as the one-sided 97.5% interval. However, it is proper to report $(-\infty, -.1)$ as the two-sided 95% interval because the direction of the effect was not known a priori.

4.4 Multivariate Aberrant Effects

If an aberrant response is defined by the joint behavior of K aspects, then the method using one aspect in Section 4.2 is not applicable. Suppose in the current section that \mathcal{A} is defined solely in terms of the K -dimensional aspect, \mathbf{Y}_i , \mathbf{y}_{Ti} , and \mathbf{y}_{Ci} , introduced in Section 3, so i exhibits an aberrant response if and only if $\mathbf{Y}_i \in \mathcal{A}$. The discussion is parallel to Section 4.2, and so will be brief.

It is possible to test hypotheses about a K -dimensional additive effect $\boldsymbol{\tau} = (\tau_1, \dots, \tau_K)^T$. Consider the following hypothesis:

$$H_0: \mathbf{y}_{Ti} = \mathbf{y}_{Ci} + \boldsymbol{\tau}_0 \text{ if either } \mathbf{y}_{Ti} \in \mathcal{A} \text{ or } \mathbf{y}_{Ci} \in \mathcal{A},$$

$$i = 1, \dots, I. \quad (7)$$

Define $\mathcal{I}_{\boldsymbol{\tau}_0} = \{i : \mathbf{Y}_i \in \mathcal{A}, \mathbf{Y}_i - Z_i\boldsymbol{\tau}_0 \in \mathcal{A}, \mathbf{Y}_i + (1 - Z_i)\boldsymbol{\tau}_0 \in \mathcal{A}\}$. In parallel with Proposition 4, part a, if the hypothesis (7) were true, then the set $\mathcal{I}_{\boldsymbol{\tau}_0}$ would be determined by \mathcal{F} . If $i \notin \mathcal{I}_{\boldsymbol{\tau}_0}$, then define $Q_{i,\boldsymbol{\tau}_0} = 0$. Let $M_{\boldsymbol{\tau}_0} = |\mathcal{I}_{\boldsymbol{\tau}_0}|$. If $i \in \mathcal{I}_{\boldsymbol{\tau}_0}$ then let $Q_{i,\boldsymbol{\tau}_0}$ be some form of unidimensional rank of the K -dimensional adjusted responses, $\mathbf{Y}_i - Z_i\boldsymbol{\tau}_0$. For instance, if higher values of each the K aspects signify greater aberration, then for $i \in \mathcal{I}_{\boldsymbol{\tau}_0}$, a simple method ranks the k th adjusted aspect, $Y_{ik} - Z_i\tau_{0k}$, from 1 to $M_{\boldsymbol{\tau}_0}$ and defines $Q_{i,\boldsymbol{\tau}_0}$ as the sum of the K ranks for the K aspects. (See Wei and Lachin 1984, sec. 3; Brunner 1991; Rosenbaum 1991, 1997; and Dawson and Lagakos 1993 for several variations on unidimensional ranks for multivariate responses, and Li, Propert, and Rosenbaum 2001, sec. 3.3, and Greevy et al. 2004 for two applications.) In parallel with Proposition 4, if (7) were true, then the ranks $\mathbf{Q}_{\boldsymbol{\tau}_0} = (Q_{1,\boldsymbol{\tau}_0}, \dots, Q_{I,\boldsymbol{\tau}_0})^T$ would be functions of \mathcal{F} , so $A_{\boldsymbol{\tau}_0} = \mathbf{Z}^T \mathbf{Q}_{\boldsymbol{\tau}_0}$ would have the null randomization distribution in Section 2.3, which is the distribution of the sum of n scores picked at random without replacement from the I scores $\mathbf{Q}_{\boldsymbol{\tau}_0} = (Q_{1,\boldsymbol{\tau}_0}, \dots, Q_{I,\boldsymbol{\tau}_0})^T$.

In short, testing one specific K -dimensional hypothesis is straightforward, in close parallel with Section 4.2. Consider the following procedure. Let \mathcal{T} be a fixed set of values of $\boldsymbol{\tau}_0$, let $\mathcal{C} \subseteq \mathcal{T}$ be the subset of values not rejected at level α , and report $(\mathcal{T}, \mathcal{C})$. In a familiar manner, the probability of falsely rejecting a true hypothesis is at most α ; that is, if the true fixed $\boldsymbol{\tau}$ is in \mathcal{T} , then $\Pr(\boldsymbol{\tau} \notin \mathcal{C} | \mathcal{F}) \leq \alpha$. With $\mathcal{T} = \{\mathbf{0}\}$, this procedure tests the hypothesis of no aberrant effects, rejecting if \mathcal{C} is empty, whereas with \mathcal{T} equal to all of K -dimensional space, \mathcal{C} is a $1 - \alpha$ confidence set, albeit one that may be difficult to describe and interpret. A useful, easy step beyond testing no aberrant effect is to use the procedure with \mathcal{T} equal to a small finite set of interesting hypotheses including $\mathbf{0}$. In a different context, Li et al. (2001, sec. 3.3) used a multivariate signed rank statistic to test $|\mathcal{T}| = 3$ hypotheses about a $K = 3$ dimensional parameter, with conclusions that are easy to describe.

5. DISCUSSION

The protocol for a clinical trial may require that, if a patient's health deteriorates in certain specific ways so that the study's treatments no longer constitute appropriate care, then that patient will be removed from the study treatments and given appropriate care. Though typically few in number, patients with

such aberrant responses may have experienced the worst clinical outcomes in the trial. The aberrations may be substantial and unacceptable deterioration in longitudinal outcomes under study, or severe side effects separate from the study outcomes, or deaths. It is natural to ask whether or not there is evidence that the treatments under study differ in their tendency to cause such aberrant responses. Because aberrant responses preclude continuation of the study treatments, the information available about patients with aberrant responses is typically different from the information available about most other patients. Because aberrant response are often defined in terms of rare but severe outcomes, the aberrant effects may be defined in terms of measures other than those used as the primary outcomes of the trial. We have proposed methods of analysis that separate aberrant effects from other effects and focus on the former. Our main recommendation is that the protocol for a clinical trial should spell out not just the definition of an aberrant response requiring removal from the study treatments, but also an analytical plan, along the lines proposed here, for judging whether one of the treatments tends to cause aberrant responses.

We developed the methods discussed here in response to the our dissatisfaction with the way in which standard methods handled aberrant responses in the AAA trial. In that trial, six of seven aberrant responses, including the only death, occurred in the placebo group, a difference that is not significant by Fisher's exact test for a 2×2 table, but which is significant when the magnitudes of the aberrations are taken into account, as was done here. Whenever a statistical method is developed in reaction to a particular data set, it is difficult to know what to make of the application of the method to that very data set; thus we are reluctant to reach conclusions about enalapril versus placebo based on our analyses. Had the same results been seen in a clinical trial whose protocol had planned for the analyses that we proposed, then these analyses would have provided fairly strong evidence that placebo had a greater tendency to cause aberrant responses.

APPENDIX: AN R FUNCTION FOR THE EXACT DISTRIBUTION

A.1 R Code

The R function `aberrant` computes the combinatorial coefficient $h_I\{a, n, \mathbf{Q}^{(I)}\}$ in (4); then the permutation distribution has $\Pr(A = a) = h_I\{a, n, \mathbf{Q}^{(I)}\} / \binom{I}{n}$.

```
> aberrant
function(I, a, n, b, d) {
  # I patients, with n treated, I-n control
  # b are not aberrant, length(d) are
  # aberrant with ranks in d
  # I = b + length(d)
  # Computes the number of treatment
  # assignments with a aberrations
  # in the treated group
  if (I==b)
    {if (a==0) out<-choose(b,n)
     else out<-0}
  else if ((a<0)|(n<0)|(I<n)) out<-0
  else {
    M<-length(d)
    out<-aberrant(I-1, a, n, b, d[1:(M-1)])
    +aberrant(I-1, a-d[M], n-1, b, d[1:(M-1)])
  }
}
```

```
}
out
}
```

A.2 Example

In Section 2.4, $I = 135$, $n = 69$, $b = 128$, and $\mathbf{d} = (1, 2, \dots, 7)$, so the number of treatment assignments $\mathbf{z} \in \Omega$ giving rise to $\mathbf{z}^T \mathbf{Q} = 2$ is:

```
>aberrant(135, 2, 69, 128, 1:7)
```

```
[1] 1.868647e+37
```

In Section 2.4, $\Pr(A \leq 2) = \Pr(A = 0) + \Pr(A = 1) + \Pr(A = 2)$ is:

```
>(aberrant(135, 0, 69, 128, 1:7)
```

```
+aberrant(135, 1, 69, 128, 1:7)
```

```
+aberrant(135, 2, 69, 128, 1:7)) /
```

```
choose(135, 69)
```

```
[1] 0.01856505
```

[Received January 2007. Revised April 2007.]

REFERENCES

- Brunner, E. (1991), "A Nonparametric Estimator of the Shift Effect for Repeated Observations," *Biometrics*, 47, 1149–1153.
- Chang, M. N., Guess, H. A., and Heyse, J. F. (1994), "Reduction in the Burden of Illness: A New Efficacy Measure for Prevention Trials," *Statistics in Medicine*, 13, 1807–1814.
- Cox, D. R. (1977), "The Role of Significance Tests" (with discussion), *Scandinavian Journal of Statistics*, 4, 49–70.
- Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman & Hall.
- Dawson, J. D., and Lagakos, S. W. (1993), "Size and Power of Two-Sample Tests of Repeated-Measures Data," *Biometrics*, 49, 1022–1035.
- Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.
- Greevy, R., Silber, J. H., Cnaan, A., and Rosenbaum, P. R. (2004), "Randomization Inference With Imperfect Compliance in the ACE-Inhibitor After Anthracycline Randomized Trial," *Journal of the American Statistical Association*, 99, 7–15.
- Lehmann, E. L. (1963), "Nonparametric Confidence Intervals for a Shift Parameter," *The Annals of Mathematical Statistics*, 34, 1507–1512.
- (1998), *Nonparametrics*, Upper Saddle River, NJ: Prentice-Hall.
- Li, Y. P., Propert, K. J., and Rosenbaum, P. R. (2001), "Balanced Risk Set Matching," *Journal of the American Statistical Association*, 96, 870–882.
- Mehrotra, D. V., Li, X., and Gilbert, P. B. (2006), "A Comparison of Eight Methods for the Dual-Endpoint Evaluation of Efficiency in a Proof-of-Concept HIV Vaccine Trial," *Biometrics*, 62, 893–900.
- Neyman, J. (1923), "On the Application of Probability Theory to Agricultural Experiments: Essay on Principles, Section 9," *Roczniki Nauk Rolniczych*, Tom X, pp. 1–51 (in Polish). Reprinted in English with discussion in *Statistical Science*, 5, 463–480.
- Rosenbaum, P. R. (1991), "Some Poset Statistics," *The Annals of Statistics*, 19, 1091–1097.
- (1997), "Signed-Rank Statistics for Coherent Predictions," *Biometrics*, 53, 556–566.
- (2001), "Effects Attributable to Treatment: Inference in Experiments and Observational Studies With a Discrete Pivot," *Biometrika*, 88, 219–231.
- (2002), *Observational Studies* (2nd ed.), New York: Springer-Verlag.
- (2005), "Heterogeneity and Causality: Unit Heterogeneity and Design Sensitivity in Observational Studies," *American Statistician*, 59, 147–152.
- Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688–701.
- Silber, J. H., Cnaan, A., Clark, B. 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.
- (2004), "Enalapril to Prevent Cardiac Function Decline in Long-Term Survivors of Pediatric Cancer Exposed to Anthracyclines," *Journal of Clinical Oncology*, 22, 820–828.
- Small, D., Ten Have, T., and Rosenbaum, P. R. (2008), "Randomization Inference in a Group-Randomized Trial of Treatments for Depression: Covariate Adjustment, Noncompliance and Quantile Effects," *Journal of the American Statistical Association*, 103, 271–279.
- 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.