JOURNAL OF BIOPHARMACEUTICAL STATISTICS, 11(4), 313–333 (2001)

EVALUATING SUBJECT-TREATMENT INTERACTION WHEN COMPARING TWO TREATMENTS

Gary L. Gadbury,[1,]* Hari K. Iyer,[2] and David B. Allison[3]

[1]Department of Mathematics and Statistics,
University of Missouri—Rolla, Rolla, Missouri 65409
[2]Department of Statistics, Colorado State University,
Fort Collins, Colorado 80523
[3]Department of Biostatistics, Section on Statistical Genetics and
Clinical Nutrition Research Center, University of
Alabama at Birmingham,
Birmingham, Alabama 35294-0022

ABSTRACT

Clinical and other studies that evaluate the effect of a treatment relative to a control often focus on estimating a mean treatment effect; however, the mean treatment effect may be misleading when the effect of the treatment varies widely across subjects. Methods are proposed to evaluate individual treatment heterogeneity (i.e., subject-treatment interaction) and its consequences in clinical experiments. The method of maximum likelihood is used to derive estimators and their properties. A bootstrap procedure that requires fewer assumptions is also presented as a small sample alternative to the maximum likelihood approach. It is shown that estimators for subject-treatment interaction are sensitive to an inestimable correlation parameter. This sensitivity is illustrated using some example data sets and using graphical plots. The practical consequence of subject-treatment interaction is that a proportion of the population may be not be responding to the treatment as indicated by the average treatment effect. Results obtained from the methods reported here can alert the practitioner to the possibility that individual treatment effects vary widely in the population and help to assess the potential consequences of this variation.

* Corresponding author. Fax: (573) 341-4741; E-mail: gadburyg@umr.edu

www.dekker.com

Applications of the proposed procedures to clinical decision making, pharmacogenetic studies, and other contexts are discussed.

Key Words: Additivity; Clinical trials; Causation; Potential response; Treatment heterogeneity

1. INTRODUCTION

A common experimental design used in clinical studies to compare two treatments is the two-sample completely randomized design. In such a setting, n_1 subjects are randomly selected to receive treatment 1 (T_1) and another n_2 subjects receive treatment 2 (T_2). We will assume that $T_1 = T$ is a new test treatment, and $T_2 = C$ is a standard or control treatment. Individual subject scores on a dependent variable (DV) are observed at some point in time. These DV could represent a change from a pretreatment baseline measurement. The two treatments are often compared by estimating an average difference in DV between T and C with respect to some population of interest (hereafter referred to as an ''average treatment effect'').

In addition to an average treatment effect, a measure of the variability in the effect of T with respect to C could help clinicians compare treatments on ''individuals'' in a study. If individual treatment variation is large with respect to the mean, the mean treatment effect could be positive and yet a nonnegligible proportion of the population could be experiencing no effect or even a negative effect. This measure of individual treatment heterogeneity or subject-treatment interaction is often overlooked when analyzing clinical results. If a measure was available, clinicians could be guided to identify a subset of the population that does not respond to treatments in a manner suggested by the estimated average treatment effect. This subset may be marked by a particular covariate that was not identified in the original design. It has been remarked (1; page 73), ''. . . if substantial variations in treatment effect from subject to subject do occur, one's understanding of the experimental situation will be very incomplete until the basis of this variation is discovered,'' and it has recently been proposed that, as a standard, inference on the average treatment effect be supplemented with inference about the variation in treatment effects (2; page 1473).

Statements about the likely variability of treatment effects can be important for several reasons. From a clinical perspective, many patients and healthcare providers might like to know not only the average effect, but also the probability of a worsening of effect. In some situations, a treatment with a lesser benefit on average might be preferred to a treatment having a superior average effect but also having a greater risk of producing a deleterious effect. Even if the effect is not plausibly deleterious, in some cases a small effect that is a ''sure thing'' may be preferred to a large average, but less dependable, effect.

There may also be forensic applications of this information. For example, a patient may experience some exacerbation in symptomology after receiving treatment and claim damages. In such a case, the probability of the treatment in question ‘‘causing’’ exacerbation would seem to be critical information. Moreover, if it were possible to estimate that probability and provide that estimate to the patient prior to initiating treatment and such information were not provided, it might put the providers or manufacturer in a vulnerable situation, namely liability.

Finally, there may be important research applications to these techniques. For example, pharmacogenomics is now a popular topic for investigation (3). Researchers are attempting to identify genetic polymorphisms that are predictive of especially good or poor drug responses. Such searches often proceed after an investigator has observed some degree of variability in the change that occurs after the application of some treatment. For example, after receiving clozapine, an antipsychotic agent, patients’ weights increase on average (4), but there is variability in the degree of weight change with some people gaining weight and some people losing weight. Some investigators have assumed that variability in *change* is equivalent to variability in *response* and examined the association of genetic polymorphisms with the degree of weight change among patients taking clozapine (5). However, this begs the question of whether there is any variability in *response* at all. The methods illustrated herein could help investigators determine the extent to which there is true variability in response before they invest great efforts in trying to determine the causes of that putative variation.

There is a fundamental issue one encounters when attempting to estimate the variation in treatment effects (i.e., subject-treatment interaction). This issue is related to what Holland (6) called ‘‘the fundamental problem of causal inference,’’ which states that at a particular moment in time and for a given subject, only a measurement of the *DV* given *T* or *C* can be observed but not both. So an individual effect of *T*, with respect to *C*, cannot be observed. Gadbury and Iyer (7) produced estimated bounds for subject-treatment interaction in a situation where a two sample design is being used and a covariate is available. They also showed, under normal distribution assumptions, that bounds for the proportion of the population experiencing an unfavorable treatment effect could be estimated. Their results depended on large sample properties of maximum likelihood estimates (MLEs).

In this paper, we extend the results of Gadbury and Iyer by assessing the ‘‘sensitivity’’ of an estimated subject-treatment interaction term to a inestimable correlation parameter. We present the technique without using covariate information and then with the use of a covariate. The results here also make use of asymptotic maximum likelihood theory and are most suitable for large samples. However, we also illustrate the bootstrap for use when the sample size is small or distributional properties of the data are not known. We conclude with a discussion.

2. DEFINING SUBJECT-TREATMENT INTERACTION

Subject-treatment interaction is present in an experimental study when the *true effect of a treatment* varies across subjects in a population. Consider a set of two potential observations (8), (X_i, Y_i), for an individual subject u_i in an investigation to compare the effect of a treatment T with respect to a control treatment C. The variable X_i is the value of the outcome on subject u_i when exposed to treatment T, and the variable Y_i is the value if exposed to treatment C. The two values are imagined to be measured at the same moment in time. In practice, only the value corresponding to the treatment actually assigned can be observed for a particular subject. Nevertheless, the two ''potential values'' help to conceptualize a true effect of treatment T with respect to treatment C on subject u_i that we define to be $D_i = X_i - Y_i$.

Suppose that potential observations (X_i, Y_i), $i = 1, 2, \dots$, are independent and identically distributed (*i.i.d.*) random variables from a bivariate distribution with mean $(\mu_X, \mu_Y)^t$ and variance matrix

$$\begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \tag{1}$$

Parameters of this bivariate distribution, with the exception of ρ_{XY}, can be estimated from the marginal distributions of X and of Y. The variability of individual treatment effects will be a function of ρ_{XY}. We assume that there is no interference between subjects (9; page 19), that is, a subject's response to a treatment does not depend on the treatment assignment outcome for other subjects in the study. The true treatment effects, D_i, have mean $\mu_D = \mu_X - \mu_Y$ and variance $\sigma_D^2 = \sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y\rho_{XY}$. Estimating μ_D is straightforward in common randomized experiments but little, if any, attention is given to evaluating σ_D^2. Subject-treatment interaction is present in the population under study when $\sigma_D^2 > 0$. When σ_D is large relative to μ_D, the mean treatment effect may not provide an adequate description of the treatment's effect on individual subjects in the population. Without loss of generality, suppose that $\mu_D > \tau$ is a beneficial treatment effect where τ is some desired threshold effect (hereafter we assume $\tau = 0$, again without loss of generality). The mean, μ_D, may be positive indicating that, on average, the treatment has a beneficial effect, and yet there may be a nonnegligible proportion, which we denote as $P_- = \Pr(D \leq 0)$, of the population that experiences an effect that is not beneficial (hereafter referred to as an unfavorable effect). A knowledge of σ_D is required to address this concern adequately. It can be shown that $\sigma_D^2 = (\sigma_X - \sigma_Y)^2 + 2\ \sigma_X\sigma_Y(1 - \rho_{XY})$, so $\sigma_D = 0$ requires that both $\sigma_X = \sigma_Y$ and $\rho_{XY} = 1$. The key issue in estimating σ_D^2 is that the correlation parameter, ρ_{XY}, is not identifiable in observed data since for each subject, either X or Y is observed but not both. Earlier results in the literature on this topic deal with testing for the presence of a subject-treatment interaction (i.e., nonadditivity) by testing for its observable consequences (10,11) or with finding transformations of the data so that subject-treatment additivity appears to hold on the transformed scale (12). However, if a ''suitable'' transformation is found, the focus then reverts to the average treatment effect on the transformed scale. This approach has at least two issues: (i) there is no information in the data for testing for unobservable consequences of subject-treatment interaction so that such interaction cannot be ruled out; and (ii) if a subset of the population experiences an unfavorable treatment effect then this subset of the population will experience this effect whether or not the data are transformed. We illustrate these ideas in the following data example (13; page 112).

Example 1

Table 1 gives the alcohol intake of 23 ''alcohol dependent'' males during a one-year period following discharge from an inpatient alcohol treatment center. Eleven individuals were randomly chosen to participate in a social skills training program (SST) plus a traditional treatment program (i.e., treatment T). The remaining 12 individuals participated in only the traditional treatment program and were thus labeled the control group (i.e., treatment C). The experiment was conducted using a two-sample completely randomized design. We have assumed that the data values, measured in centiliters (cl), accurately represent the alcohol intake for the one-year period and that treatment compliance was not an issue.

A point estimate of the difference in mean alcohol intake between the two groups, μ_D, is equal to -456 cl, and a two-sample t-distribution based 95% confidence interval for this mean difference is (-694 cl, -218 cl). These results provide some evidence in favor of the SST program in reducing average alcohol consumption in alcohol-dependent males.

Still, we have not considered an important aspect of the treatment's effect on the *individuals* in the study. Observe that subject 1 in the SST group had a one-year alcohol intake of 874 cl. We cannot know what that particular subject's

Table 1. Alcohol Intake for 1 Year (Centiliter of Pure Alcohol, cl)

Subject	SST	Subject	Control
1	874	12	1,042
2	389	13	1,617
3	612	14	1,180
4	798	15	973
5	1,152	16	1,552
6	893	17	1,251
7	541	18	1,151
8	741	19	1,511
9	1,064	20	728
10	862	21	1,079
11	213	22	951
		23	1,319

alcohol intake would have been if he had been assigned to the control group instead, and similarly for all subjects in the study. The true treatment effect, D_i $i = 1, \ldots, 23$, cannot be observed. For this reason it is not possible to directly estimate σ_D (or P_-) from observable data. However, it is possible to assess the sensitivity of an estimate to varying ρ_{XY}.

3. EVALUATING σ_D AND $P_- = \Pr(D \leq 0)$

Suppose that potential observations (X, Y) are bivariate normal (possibly after a suitable transformation) with mean vector $(\mu_X, \mu_Y)^t$ and covariance matrix given by Eq. (1). Let X_i, $i = l, \ldots, n_1$, denote the observed values for the n_1 subjects assigned to the treatment group in a two sample completely randomized design. Likewise, let Y_j, $j = 1, \ldots, n_2$, denote the observed values for the n_2 subjects assigned to the control group. The likelihood function of observed data is of the form, $\Pi_{i=1}^{n_1} f(x_i) \Pi_{j=1}^{n_2} f(y_j)$.

For a given value of ρ_{XY}, the maximum likelihood estimator (MLE) for σ_D^2 is given by,

$$\hat{\sigma}_D^2 = \hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_X\hat{\sigma}_Y\rho_{XY} \tag{2}$$

where

$$\begin{aligned} \hat{\sigma}_X^2 &= s_X^2 = (1/n_1)\sum_{i=1}^{n_1} (x_i - \bar{x})^2 \\ \hat{\sigma}_Y^2 &= s_Y^2 = (1/n_2)\sum_{j=1}^{n_2} (y_j - \bar{y})^2 \end{aligned} \tag{3}$$

and $\bar{x}$ and $\bar{y}$ are the arithmetic sample means of observed X and Y, respectively. Furthermore, the large sample distribution of $\hat{\sigma}_D^2$ is approximately normal with mean σ_D^2 and variance,

$$\operatorname{Var}(\hat{\sigma}_D^2) = 2\left\{\frac{\sigma_X^2}{n_1}(\sigma_X - \rho_{XY}\sigma_Y)^2 + \frac{\sigma_Y^2}{n_2}(\sigma_Y - \rho_{XY}\sigma_X)^2\right\} \tag{4}$$

The derivation of Eq. (4) is outlined in the Appendix. From this result one can assess the sensitivity of $\hat{\sigma}_D$, and corresponding large sample confidence bands for σ_D, to varying values of ρ_{XY} between -1 and 1.

Assuming, again without loss of generality, that $\mu_D > 0$, then the probability of an unfavorable treatment effect is given by

$$P_- = \Phi(-\mu_D/\sigma_D)$$

where $\Phi(a)$ is the cumulative standard normal distribution function evaluated at a. The MLE of μ_D is $\hat{\mu}_D = \bar{x} - \bar{y}$, so for a given value of ρ_{XY}, the maximum likelihood estimator for P_- is

$$\hat{P}_- = \Phi(-\hat{\mu}_D/\hat{\sigma}_D)$$

where $\hat{\sigma}_D$ is the square root of $\hat{\sigma}^2_D$, given in Eq. (2). The large sample distribution of $\hat{P}_-$ is approximately normal with mean P_- and variance,

$$\mathrm{Var}(\hat{P}_-) = \frac{(\phi(-\mu_D/\sigma_D))^2}{\sigma^2_D}\left\{\mathrm{Var}(\hat{\mu}_D) + \frac{\mu^2_D\ \mathrm{Var}(\hat{\sigma}^2_D)}{4\sigma^4_D}\right\} \tag{5}$$

where $\phi(a)$ is the standard normal density evaluated at a, $\mathrm{Var}(\hat{\sigma}^2_D)$ is given in Eq. (4), and $\mathrm{Var}(\hat{\mu}_D) = \sigma^2_X/n_1 + \sigma^2_Y/n_2$ [see Appendix for the derivation of Eq. (5)].

A Return to Example 1

Though the Example 1 data came from a small data set, we use it to illustrate the results from above. Inference on the average treatment effect suggested that the SST treatment was beneficial. The following results give the investigator added information regarding the proportion of population individuals that benefit from the SST treatment. The relevant MLEs are as follows:

$$\hat{\mu}_D = -456, \quad \hat{\sigma}_X = 268, \quad \hat{\sigma}_Y = 257$$

The estimated standard deviation of treatment effects, $\hat{\sigma}_D$, ranges from a high of 524.8 down to 11.7 as ρ_{XY} varies from -1 to 1. Figure 1 shows the sensitivity of $\hat{\sigma}_D$ to varying ρ_{XY} along with large sample 95% confidence bands for σ_D. The figure suggests that for most values of ρ_{XY}, there is some subject-treatment interaction (i.e., $\sigma_D > 0$). But is it enough to indicate that some subset of the population would be better off or at least as well off with only the traditional treatment rather than the SST treatment?

The MLE for P_- ranges from a high of 0.192 down to zero as ρ_{XY} varies from -1 to 1. Figure 2 shows the MLE values for varying ρ_{XY} in addition to large sample 95% confidence bands for P_-. The lower confidence band suggests that for most values of ρ_{XY}, there is insufficient evidence in the data to suggest that a positive proportion of the population will experience an unfavorable effect due to the SST treatment program. In fact, when ρ_{XY} exceeds 0.80, the upper confidence limit for P_- is less than 0.018 indicating that it is unlikely that an individual would experience an unfavorable effect of the SST treatment.

On the other hand, if $\rho_{XY} = -1$, then a large sample 95% confidence interval for P_- is (0.061,0.324). Although $\rho_{XY} = -1$ is theoretically possible, in many real applications one may believe that ρ_{XY} is actually closer to 1. In such cases, the sensitivity can be restricted to a narrower range. In fact, when $\rho_{XY} = 1$, the magnitude of σ^2_D depends on the difference between σ_X and σ_Y, and these two parameters can be estimated from observed data. Since, however, one will never know if ρ_{XY} actually equals one, the sensitivity of σ_D to moderate values of ρ_{XY} may be of interest.

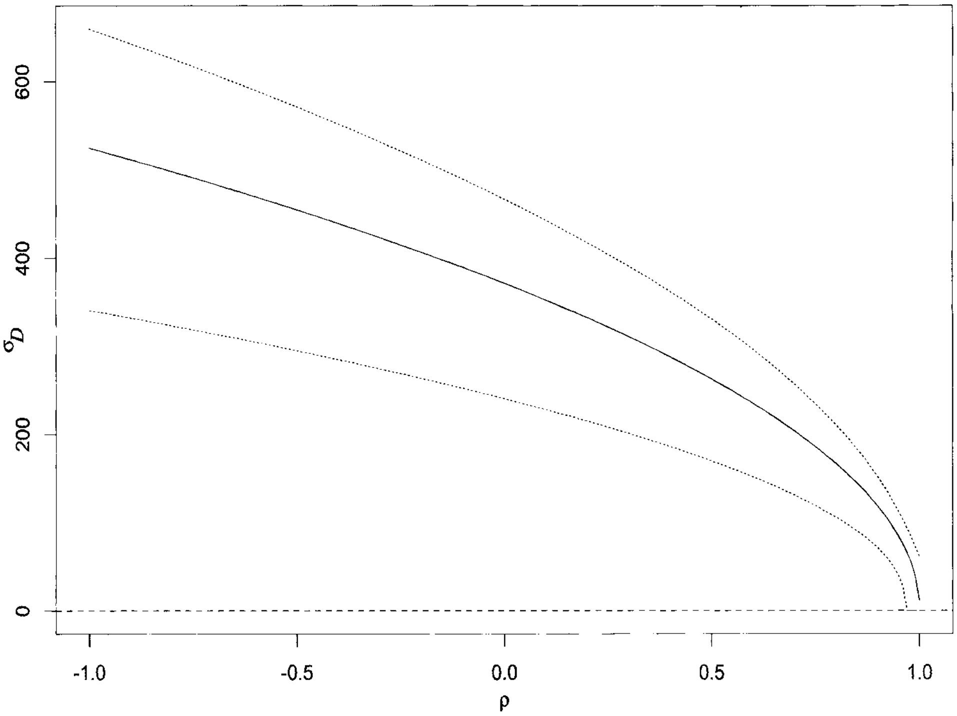

Figure 1. Sensitivity of estimated σ_D to varying $\rho_{XY} = \rho$ for Example 1. The solid line is the MLE for σ_D, and the dotted lines are 95% confidence bands.

Caution must be exercised when using the confidence bands with such small data sets. The confidence bands rely on asymptotic normal theory of maximum likelihood estimators, and they will be more accurate with much larger data sets. Later, in Section 5, we present a bootstrap procedure as an alternative to the maximum likelihood method, and we illustrate it using these data. In the next section we consider the role of a covariate when evaluating subject-treatment interaction and its consequences.

4. THE ROLE OF A COVARIATE

Suppose now that a covariate, Z, is observable on all subjects in the sample. As usual, the covariate is assumed to have been observed before application of treatment, or to not be influenced by the treatment. We now seek to estimate P_- for any given subpopulation of subjects with a specified value of Z. We denote this proportion as $P_{-\cdot Z}$ and we assess its sensitivity to an inestimable partial correlation parameter. Lower and upper bounds for the unconditional P_- have been derived (making use of covariate information) along with their corresponding MLEs and were reported in an earlier work (7).

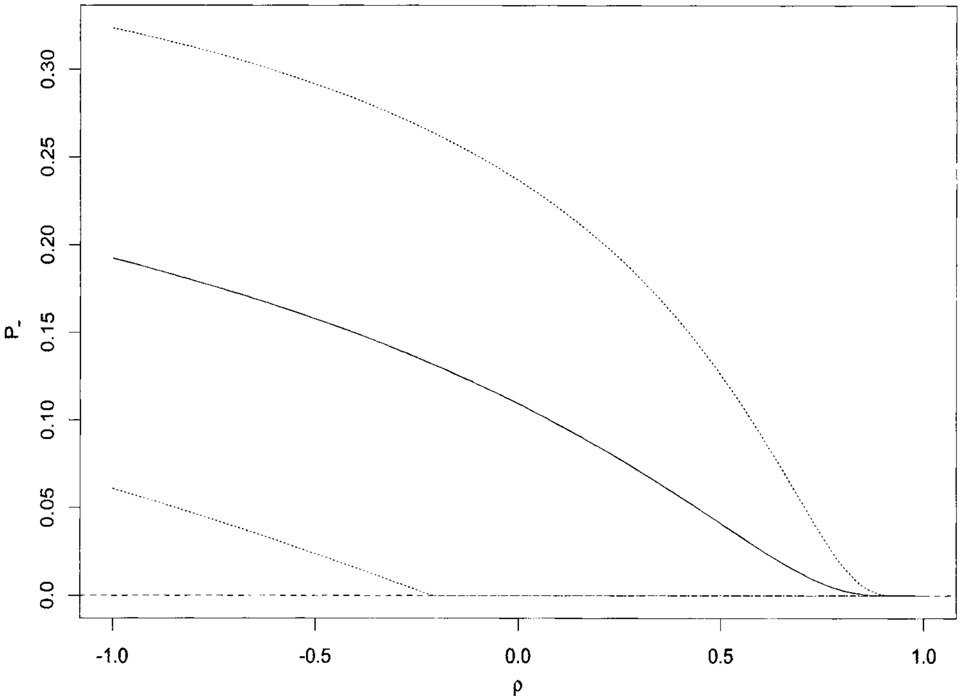

Figure 2. Sensitivity of estimated P_- to varying $\rho_{XY} = \rho$ for Example 1. The solid line is the MLE for P_-, and the dotted lines are 95% confidence bands.

The population of potential observations may now be viewed as a trivariate population, which we assume to be normal (possibly after suitable transformations), represented by the random vector (X, Y, Z). Let this random vector have mean $(\mu_X, \mu_Y, \mu_Z)^t$ and variance matrix

$$\begin{pmatrix} \sigma_X^2 & \rho_{XY}\sigma_X\sigma_Y & \rho_{XZ}\sigma_X\sigma_Z \\ \rho_{XY}\sigma_X\sigma_Y & \sigma_Y^2 & \rho_{YZ}\sigma_Y\sigma_Z \\ \rho_{XZ}\sigma_X\sigma_Z & \rho_{YZ}\sigma_Y\sigma_Z & \sigma_Z^2 \end{pmatrix} \tag{6}$$

The parameters of this distribution, except ρ_{XY}, can be estimated from the marginal distributions of X, Y, and Z, and from the bivariate distribution of (X, Z), and of (Y, Z). The population linear regression functions relating the conditional means of X and Y given $Z = z_0$ are, respectively,

$$\mu_{X \cdot Z=z_0} = \mu_X + \beta_X(z_0 - \mu_Z)$$

$$\mu_{Y \cdot Z=z_0} = \mu_Y + \beta_Y(z_0 - \mu_Z)$$

where $\beta_X = \rho_{XZ}\sigma_X/\sigma_Z$ and $\beta_Y = \rho_{YZ}\sigma_Y/\sigma_Z$. There is no subject-treatment interaction, i.e., $\sigma_D^2 = 0$, if and only if the following three conditions are satisfied: (i)

$\beta_X = \beta_Y$, (ii) $\sigma_{X\cdot Z} = \sigma_{Y\cdot Z}$, and (iii) $\rho_{XY\cdot Z} = 1$, where $\sigma_{X\cdot Z}$ and $\sigma_{Y\cdot Z}$ are conditional standard deviations of X and Y, respectively, given Z and $\rho_{XY\cdot Z}$ is the partial correlation of X and Y given Z. Proof of this assertion follows from the identity

$$\sigma_D^2 = (\sigma_{X\cdot Z} - \sigma_{Y\cdot Z})^2 + 2\sigma_{X\cdot Z}\sigma_{Y\cdot Z}(1 - \rho_{XY\cdot Z}) + (\beta_X - \beta_Y)^2\sigma_Z^2 \tag{7}$$

Only conditions (i) and (ii) above can be tested using observable data.

If observed data provide evidence that $\beta_X \neq \beta_Y$, one might argue that this information could be used to predict positive (or negative) treatment effects on the basis of covariate values. What is actually predicted in such a case is the mean treatment effect conditioned on a covariate value. For a given covariate value, say $Z = z_0$, there is a subpopulation of *individual treatment effects* for that given value of Z that is normal with mean equal to $\mu_{D\cdot Z=Z_0}$ and variance equal to $\sigma^2_{D\cdot Z=z_0} = \sigma^2_{D\cdot Z}$ where,

$$\mu_{D\cdot Z=Z_0} = \mu_X - \mu_Y + (\beta_X - \beta_Y)(z_0 - \mu_Z)$$

$$\sigma^2_{D\cdot Z} = \sigma^2_{X\cdot Z} + \sigma^2_{Y\cdot Z} - 2\sigma_{X\cdot Z}\sigma_{Y\cdot Z}\rho_{XY\cdot Z}$$

The partial correlation, $\rho_{XY\cdot Z}$, cannot be estimated from observed data, but it must lie in the interval $(-1, 1)$.

Let (X_i, Z_{1i}), $i = 1, \ldots, n_1$, be observable values of the test treatment variable and the value of the covariate for the n_1 subjects assigned to the treatment group. Likewise, let (Y_j, Z_{2j}), $j = 1, \ldots, n_2$, be observable values for the n_2 subjects assigned to the control group. The likelihood function of observed data is of the form

$$\prod_{i=1}^{n_1} f(x_i, z_{1i}) \prod_{j=1}^{n_2} f(y_j, z_{2j}) \tag{8}$$

For a given $\rho_{XY\cdot Z}$, the MLE of $\sigma^2_{D\cdot Z=z_0}$ is given by

$$\hat{\sigma}^2_{D\cdot Z} = \hat{\sigma}^2_{X\cdot Z} + \hat{\sigma}^2_{Y\cdot Z} - 2\,\hat{\sigma}_{X\cdot Z}\hat{\sigma}_{Y\cdot Z}\rho_{XY\cdot Z}$$

with

$$\hat{\sigma}^2_{X\cdot Z} = s^2_{X\cdot Z} = s^2_X(1 - r^2_{XZ}),\ \hat{\sigma}^2_{Y\cdot Z} = s^2_{Y\cdot Z} = s^2_Y(1 - r^2_{YZ})$$

where s^2_X and s^2_Y are given in Eq. (3), and r_{XZ} and r_{YZ} are the usual sample correlation coefficients. The large sample distribution of $\hat{\sigma}^2_{D\cdot Z}$ is normal with mean $\sigma^2_{D\cdot Z}$ and variance

$$\mathrm{Var}(\hat{\sigma}^2_{D\cdot Z}) = 2\left\{\frac{\sigma^2_{X\cdot Z}}{n_1}(\sigma_{X\cdot Z} - \rho_{XY\cdot Z}\sigma_{Y\cdot Z})^2 + \frac{\sigma^2_{Y\cdot Z}}{n_2}(\sigma_{Y\cdot Z} - \rho_{XY\cdot Z}\sigma_{X\cdot Z})^2\right\}$$

The MLE of $\mu_{D\cdot Z=z_0}$ is $\hat{\mu}_{D\cdot Z} = \bar{x} - \bar{y} + b_X(z_0 - \bar{z}_1) - b_Y(z_0 - \bar{z}_2)$, where $\hat{\beta}_X = b_X = s_X r_{XZ}/s_{Z_1}$, $\hat{\beta}_Y = b_Y = s_Y r_{YZ}/s_{Z_2}$, $\bar{x}$ and $\bar{z}_1$ are observed sample means of the n_1 individuals in the treatment group and similarly for $\bar{y}$ and $\bar{z}_2$, s_{Z_1} is the sample standard deviation of covariate values (divisor n_1) for the n_1 observations in the treatment group and similarly for s_{Z_2}. The estimator $\hat{\mu}_{D\cdot Z}$ is asymptotically normal with mean $\mu_{D\cdot Z=z_0}$ and variance

$$\mathrm{Var}(\hat{\mu}_{D\cdot z}) = (\sigma^2_{X\cdot Z}/n_1 + \sigma^2_{Y\cdot Z}/n_2)\left(1 + \frac{(z_0 - \mu_Z)^2}{\sigma^2_Z}\right)$$

The above equations can be derived using results in Lord (14) who provided MLEs and large sample variances of the eight individual parameters in Eq. (8). The derivation again uses properties of MLE's and is similar to the derivation of results from Section 3, shown in the Appendix.

Assuming, without loss of generality, that for a given z_0, $\mu_{D\cdot Z=z0} > 0$, then the probability that an individual experiences a negative effect is $P_{-\cdot Z=z0}$ where

$$P_{-\cdot Z=z_0} = \Phi(-\mu_{D\cdot Z=z_0}/\sigma_{D\cdot Z})$$

For a given $\rho_{XY\cdot Z}$, the MLE of $P_{-\cdot Z=z0}$ is given by

$$\hat{P}_{-\cdot Z} = \Phi(-\hat{\mu}_{D\cdot Z}/\hat{\sigma}_{D\cdot Z})$$

which is asymptotically normal with mean $P_{-\cdot Z=z_0}$ and variance

$$\mathrm{Var}(\hat{P}_{-\cdot Z}) = \frac{(\phi(-\mu_{D\cdot Z}/\sigma_{D\cdot Z}))^2}{\sigma^2_{D\cdot Z}}\left\{\mathrm{Var}(\hat{\mu}_{D\cdot Z}) + \frac{\mu^2_{D\cdot Z}\,\mathrm{Var}(\hat{\sigma}^2_{D\cdot Z})}{4\sigma^4_{D\cdot Z}}\right\}$$

Results in this section are particularly useful when the slopes of the two regression lines relating X and Z and relating Y and Z appear unequal. Sensitivity of $\hat{\sigma}_{D\cdot Z}$ and of $\hat{P}_{-\cdot Z}$ at a given value of Z can be assessed for varying $\rho_{XY\cdot Z}$. We illustrate this using a well known small data set (15; page 552).

Example 2

Again, we use a small data set for illustration though results from these methods will be more accurate for larger data sets. The example data are shown in Table 2. A baseline seated systolic blood pressure, Z, was recorded for 21 male subjects. The subjects were randomized into two groups so that 10 subjects received a calcium supplement (the treatment), and the other 11 subjects received a placebo. After a period of 12 weeks, the seated systolic blood pressure was again recorded R, and a change from baseline $C = R - Z$ was computed. The experiment was double-blind.

Define a treatment indicator variable W that is equal to 1 for subjects receiving the calcium supplement and equal to zero otherwise. A linear model

$$C = \beta_0 + \beta_1 W + \beta_2 Z + \beta_3(Z \times W) + \varepsilon$$

was fit to the data where ε is an error term assumed to be from a standard normal distribution. The estimated coefficients from the model are

Table 2. Blood Pressure Measurements

Subject	Z	R	C
Treatment			
1	107	100	−7
2	110	114	4
3	123	105	−18
4	129	112	−17
5	112	115	3
6	111	116	5
7	107	106	−1
8	112	102	−10
9	136	125	−11
10	102	104	2
Control			
11	123	124	1
12	109	97	−12
13	112	113	1
14	102	105	3
15	98	95	−3
16	114	119	5
17	119	114	−5
18	112	114	2
19	110	121	11
20	117	118	1
21	130	133	3

Z is a baseline measure. *R* is the blood pressure after 12 weeks. Changes from baseline are $C = R - Z$.

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (-8.002, 68.122, 0.076, -0.643)$$

Recall that the potential outcome variables are X and Y. The estimated model relating mean change from baseline for subjects on the calcium treatment is $\hat{\mu}_X = 60.120 - 0.567Z$ and the corresponding model for the placebo group is $\hat{\mu}_Y = -8.002 + 0.076Z$. The estimated mean treatment effect is expressed as $\hat{\mu}_D = \hat{\mu}_X - \hat{\mu}_Y = 68.122 - 0.643Z$, which implies that the estimated mean treatment effect depends on the baseline blood pressure Z. A plot of the observed data and fitted regression lines is shown in Figure 3. The figure shows that subjects on the calcium treatment experience, on average, a greater decrease in blood pressure for larger values of baseline blood pressure. There was little change in blood pressure for subjects on the placebo. The treatment by baseline interaction is apparent from the unequal slopes of the fitted regression lines for each group. The linear model allows estimation of a mean treatment effect at any given value of Z, but it does not provide any indication of individual variability of treatment effects at that value of Z. We proceed with this example using the techniques in this section.

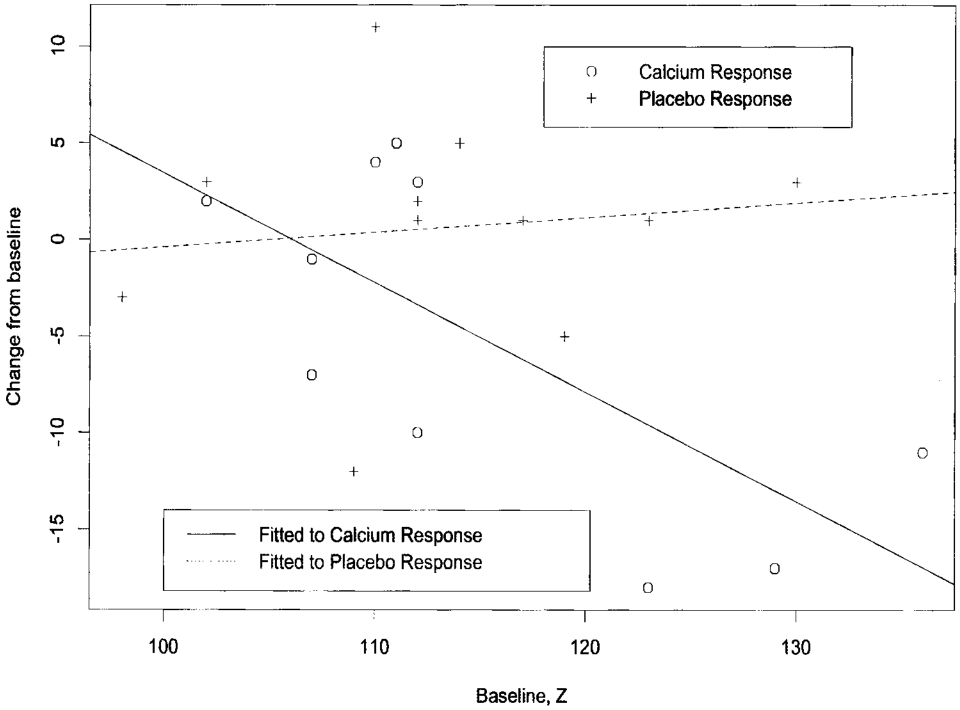

Figure 3. Plot of changes from baseline versus baseline blood pressure for Example 2.

The standard deviation of treatment effects, $\sigma_{D\cdot Z}$, does not depend on values of Z (under normality). Figure 4 shows the sensitivity of estimated $\sigma_{D\cdot Z}$ to varying partial correlation of X and Y given Z. The plot suggests that for most values of $\rho_{XY\cdot Z}$ between -1 and 1 there is evidence of subject-treatment interaction, that is, $\sigma_{D\cdot Z} > 0$ in the subpopulations defined by given values of Z.

The average baseline blood pressure is $\bar{z} = 114$. An estimated average treatment effect, at $\bar{z} = 114$, has point estimate $\hat{\mu}_{D\cdot Z=114} = -5.21$ and a 95% confidence interval $(-11.61, 0.68)$. This suggests that calcium may be marginally effective, *on average*, in reducing blood pressure of individuals with a baseline blood pressure of 114 (a one tailed P-value testing an alternative *Ha*: $\mu_{D\cdot Z=114} < 0$ is 0.039). Figure 5 shows the sensitivity of estimated $P_{-\cdot Z}$ to varying $\rho_{XY\cdot Z}$ between -1 and 1. Based on this figure it appears that, if $\rho_{XY\cdot Z}$ is positive, the data do not suggest that there is a positive proportion of the subpopulation at $Z = 114$ that would experience an increase in blood pressure due to the calcium treatment. Yet the upper confidence band indicates that the proportion could be high. For example, if $\rho_{XY\cdot Z} = 0.5$, $P_{-\cdot Z}$ is between 0 and 0.42 with 95% confidence. The confidence bands are wide due to the small sample size.

The mean treatment effect for a subpopulation with $Z = 130$ is estimated to be in the interval $(-25.06, -5.88)$ with 95% confidence. Furthermore, individuals with this high baseline blood pressure are also more likely to benefit from the

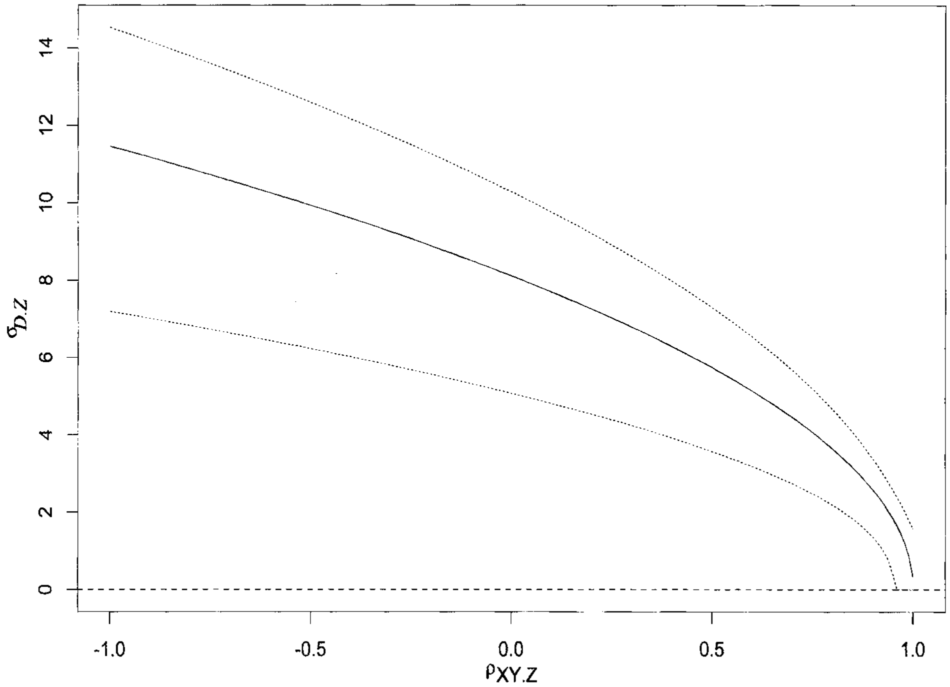

Figure 4. Sensitivity of estimated $\sigma_{D\cdot Z}$ to varying $\rho_{XY\cdot Z}$ for Example 2. The solid line is the MLE for $\sigma_{D\cdot Z}$, and the dotted lines are 95% confidence bands.

calcium supplements. This is shown in Figure 6. If $\rho_{XY\cdot Z} = 0.5$, then on this graph $P_{-\cdot Z}$ is between 0 and 0.02 with 95% confidence. Even in the worst case when $\rho_{XY\cdot Z} = -1$, the lower confidence limit for $P_{-\cdot Z}$ is still zero. So Figure 6 provides some indication that if a person's blood pressure is high, then not only will the population average blood pressure decrease, but most *individuals* will benefit as well. This analysis can be repeated for any subpopulation of interest defined by a value of Z.

A final note regarding this example is that as ρ_{XZ} and ρ_{YZ} approach 1, then $\sigma_{D\cdot Z}$ goes to zero. This does not mean there is no subject-treatment interaction present in the population, but it does mean that any subject-treatment interaction can be explained by the covariate Z. This fact highlights the need to find covariates that are good predictors of outcomes, and the sample correlation coefficients provide some indication of this predictive capability. In this example, $\hat{\rho}_{XZ} = 0.602$ and $\hat{\rho}_{YZ} = 0.857$.

5. AN APPROACH FOR SMALL SAMPLES

The methods described thus far entail normal distribution theory and large sample confidence intervals. In situations when the distribution of data is unknown

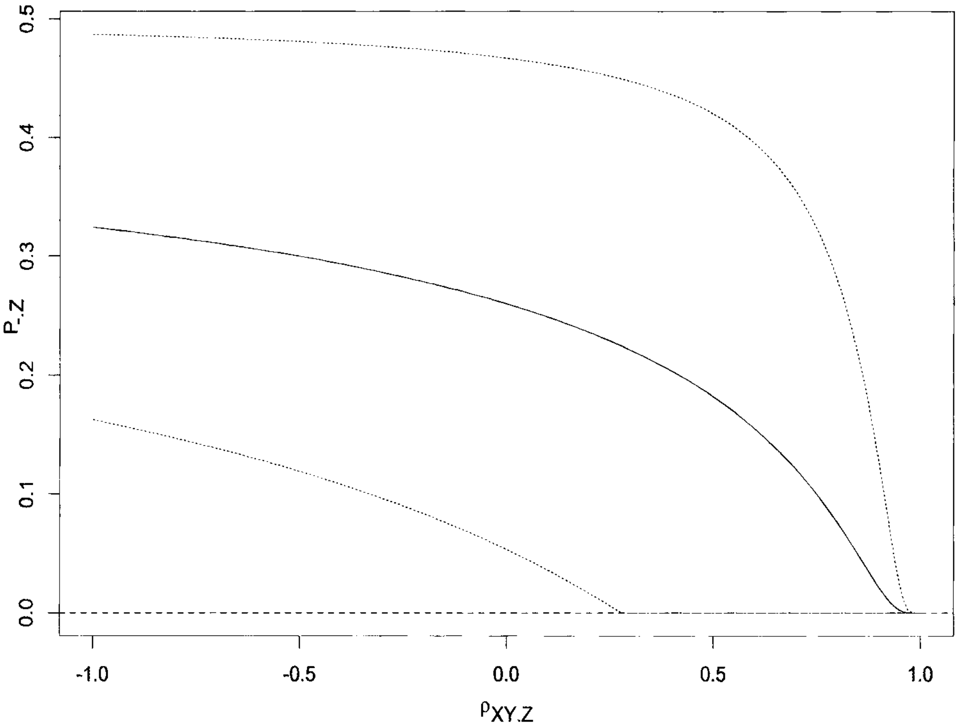

Figure 5. Sensitivity of estimated $P_{-\cdot Z}$ evaluated at $Z = \bar{z} = 114$ to varying $\rho_{XY\cdot Z}$ for Example 2. The solid line is the MLE for $P_{-\cdot Z}$, and the dotted lines are 95% confidence bands.

or sample sizes are small (as in the examples we used), one could use a bootstrap procedure. Details regarding the bootstrap can be found in Efron and Tibshirani (16). We highlight the key points below in the context of a two sample design without a covariate.

We assume that potential observations are, again, a random sample from a larger bivariate population (not necessarily normal). After treatment assignment, we observe n_1 values in response to T and n_2 in response to C.

For a given correlation, ρ_{XY}, the point estimator of σ_D^2 is given by,

$$\hat{\sigma}_D^2 = \hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_X\hat{\sigma}_Y\rho_{XY}$$

where, in this case, we use the unbiased estimators of variance. That is,

$$\hat{\sigma}_X^2 = (1/(n_1 - 1))\sum_{i=1}^{n_1} (x_i - \bar{x})^2$$

$$\hat{\sigma}_Y^2 = (1/(n_2 - 1))\sum_{j=1}^{n_2} (y_i - \bar{y})^2$$

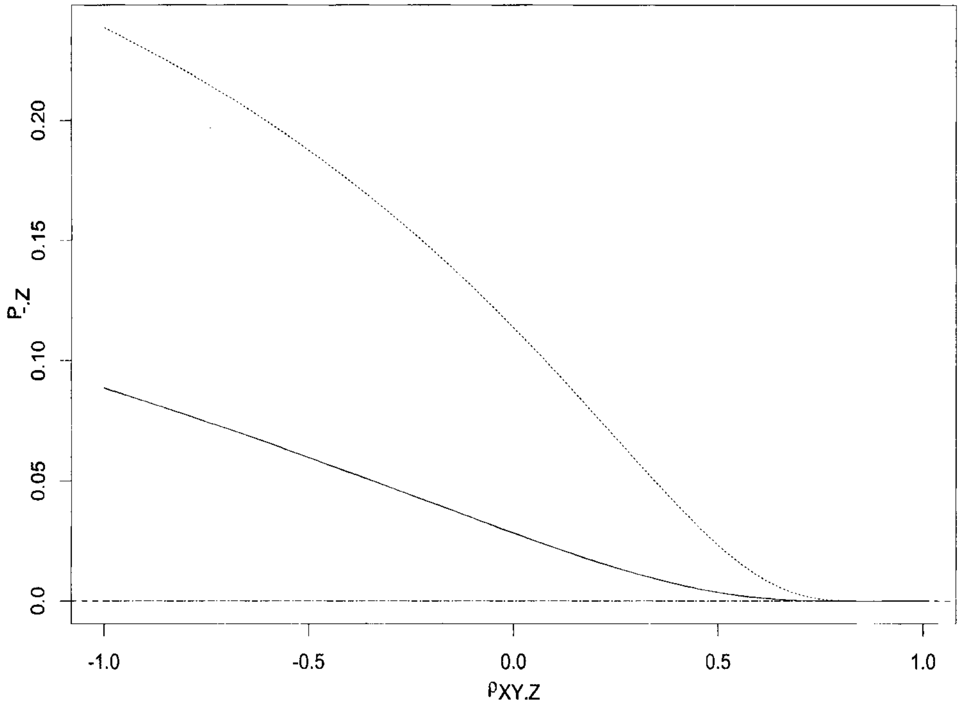

Figure 6. Sensitivity of estimated P_{-Z} evaluated at $Z = 130$ to varying $\rho_{XY \cdot Z}$ for Example 2. The solid line is the MLE for P_{-Z}, and the dotted lines are 95% confidence bands.

A bootstrap sample of the treatment group is drawn by resampling n_1 observed values with replacement from the actual n_1 outcomes from treatment T. A bootstrap sample of the control group is similarly obtained. Denote a bootstrap sample from the test treatment group as $(x_1^*, x_2^*, \ldots, x_{n1}^*)$. The bootstrap estimate of σ_X^2 is given by

$$s_X^{*2} = \frac{n_1}{n_1 - 1}(1/(n_1 - 1))\sum_{i=1}^{n_1} (x_i^* - \bar{x}^*)^2$$

where $\bar{x}^*$ is the mean of the bootstrap sample. The usual sample variance of a bootstrap sample will be biased low, and so the fraction $(n_1/(n_1 - 1))$ was included to correct for this. Similarly

$$s_Y^{*2} = \frac{n_2}{n_2 - 1}(1/(n_2 - 1))\sum_{i=1}^{n_2} (y_i^* - \bar{y}^*)^2$$

is the variance of the bootstrap sample from the control group. For a given ρ_{XY}, a bootstrap estimate of σ_D^2 is given by

$$\hat{\sigma}_D^{*2} = \hat{\sigma}_X^{*2} + \hat{\sigma}_Y^{*2} - 2\hat{\sigma}_X^*\hat{\sigma}_Y^*\rho_{XY}$$

For each ρ_{XY}, B bootstrap samples can be drawn and a value of $\hat{\sigma}_D^{*2}$ can be calculated. Let F_ρ^* be the bootstrap distribution of values of $\hat{\sigma}_D^{*2}$ for a given ρ_{XY}. Bootstrap $(1 - \alpha)100\%$ confidence intervals for σ_D^2 are the $\alpha/2$ and $1 - \alpha/2$ quantiles of F_ρ^*. These are percentile confidence intervals (16). When this is done for values of ρ_{XY} in the interval $(-1, 1)$ one obtains $(1 - \alpha)100\%$ confidence bands for σ_D^2. Since percentile intervals are transformation respecting, corresponding confidence bands for σ_D can be computed using a square root transformation on the values comprising the confidence bands for σ_D^2.

We used this method with the data in Table 1, and the results are shown in Figure 7. Results are similar to those of Example 1 where MLEs were used. However, unlike the confidence bands for the MLE, the sample estimate of σ_D is not always centered in the bootstrap confidence bands. This is not unusual for bootstrap percentile confidence intervals. There are many other bootstrap methods for obtaining confidence intervals, and their strengths and weaknesses depend on the particular application (16).

Estimating P_- would require either an assumed distribution for treatment effects (as was made when using MLEs), or a method to bound probabilities such as Chebyshev's inequality, the latter being possibly too conservative for many applications. To continue with the Example 1 data, recall that μ_D had a t-distribu-

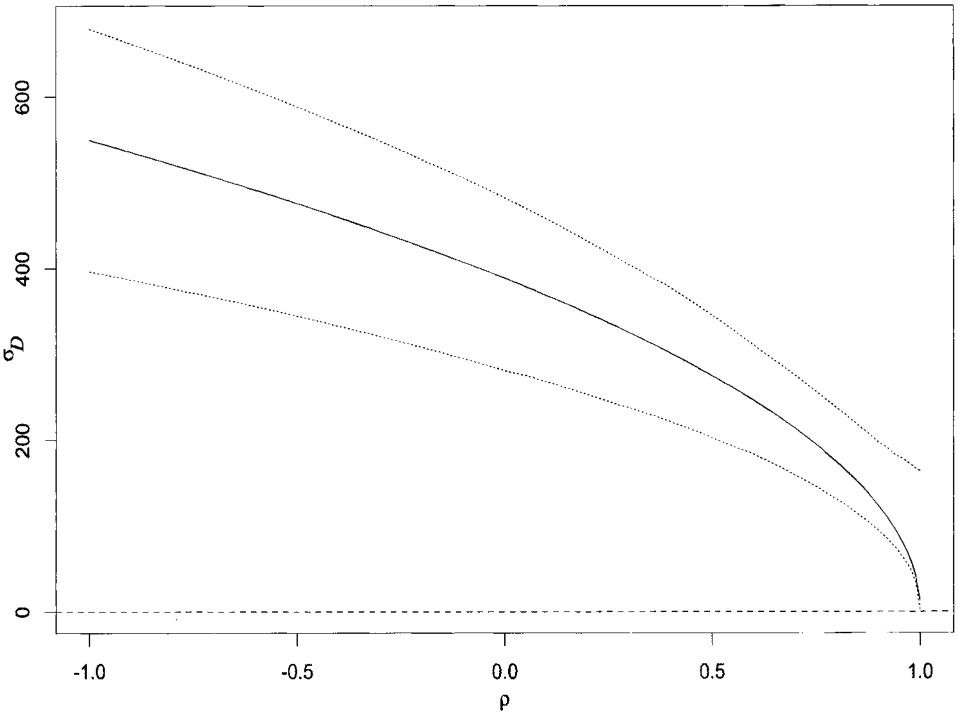

Figure 7. Bootstrap 95% confidence bands (dotted lines) for σ_D for varying $\rho_{XY} = \rho$ using data in Table 1. The middle solid line is the sample point estimate given in Section 5.

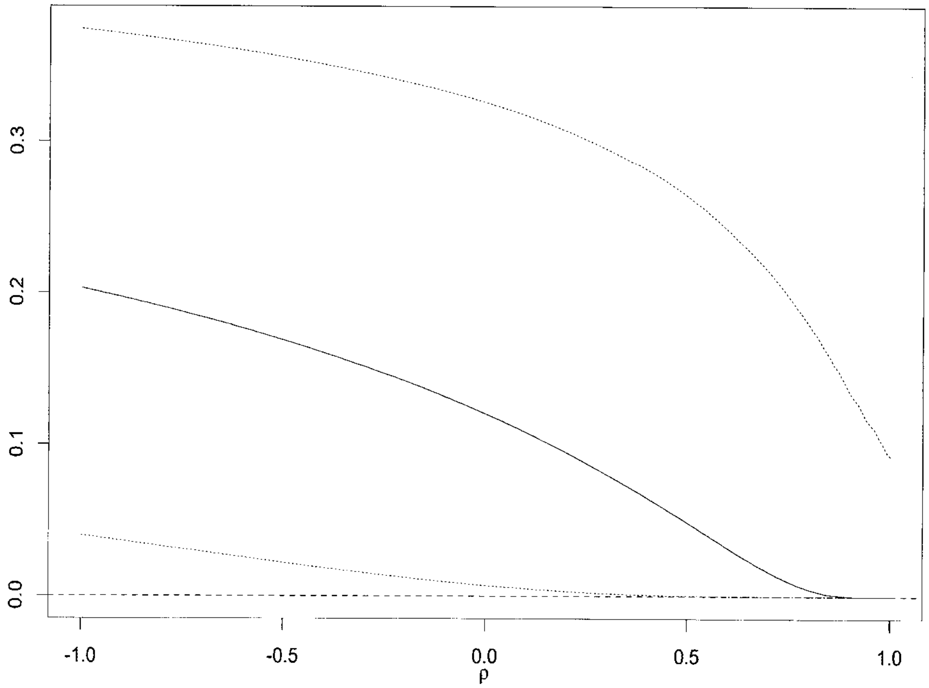

Figure 8. 90% ''small sample'' confidence bands (dotted lines) for P_- for varying $\rho_{XY} = \rho$ using data in Table 1. The middle solid line is the MLE for P_-.

tion based 95% confidence interval (−694, −218). Also, from above we have 95% bootstrap confidence bounds for σ_D at each value of ρ_{XY}. The two sets together can provide conservative 90% confidence bands for P_-. The result is shown in Figure 8.

The confidence bands are wide but generally follow a similar pattern to Figure 2. One exception is that the lower bound is ''range respecting'' meaning that the bound is never negative since the parameter it estimates (i.e., a probability) is never negative (this was not true for the confidence bands given in Sections 3 and 4). A second difference is that the upper confidence band is larger than that of Figure 2 as ρ_{XY} approaches one. This may reflect uncertainty due to the small sample size and/or the conservative nature of the joint confidence region for μ_D and σ_D since the joint region was assumed to be rectangular. The exact joint confidence region will likely be more complex than a simple rectangle. This is a subject for further research.

6. DISCUSSION

In this paper we presented methods to evaluate subject-treatment interaction and its consequences using a two-sample randomized design, and we discussed possible applications for the methods. Throughout, we have assumed no measurement errors and we have assumed treatment compliance. The role of measurement errors and treatment compliance will be discussed in the future work.

We also noted that when the sample size is small, caution must be exercised when interpreting the confidence bands obtained from maximum likelihood theory. The bootstrap bounds may be more accurate in such cases. A subject for further research is to compare confidence bands obtained using properties of maximum likelihood estimators with exact confidence bands using normal distribution theory. The exact confidence bands may be obtained from simultaneous joint confidence regions for σ_X^2, σ_Y^2, and μ_D. For a fixed ρ_{XY}, minimizing and maximizing the expressions for σ_D^2 and P_- would produce confidence intervals for these parameters at that value of ρ_{XY}. Moreover, when a covariate is available, only the conditional distributions of X given Z and of Y given Z would need to be assumed normal. This is similar to an assumption made when conducting inference using linear regression models.

Since the primary issue in estimating subject-treatment interaction has been the fact that individual treatment effects cannot be observed at a single point in time, a natural question arises about the use of crossover designs to circumvent this issue. In such a design, subjects are randomly assigned to a ''treatment sequence.'' A subject receives both treatments at different points in time separated by a washout period. So an individual subject's outcome for each treatment can be observed, and an ''individual treatment effect'' can be computed. In a two-period crossover design, even if we can safely assume absence of carryover effects, there are four potential observations, $(X^{(j)}, Y^{(j)})$ where $j = 1, 2$ denotes the time period at which one would measure an observation. Only one of the two pairs, $(X^{(1)}, Y^{(2)})$ or $(X^{(2)}, Y^{(1)})$, can be observed for an individual depending on which sequence of treatments the individual received. Evaluating subject-treatment interaction in various crossover designs is a subject for further research. Some results for a two period balanced crossover design using potential outcomes are in Gadbury (17).

Finally, conclusions based on the data alone may not be definitive enough (usually due to small sample sizes involved in many studies) and this knowledge is often combined with subject matter knowledge relevant to the particular application. For example, the practical interpretation of an ''unfavorable treatment effect'' is disease/disorder specific. But results obtained from the methods reported here can not only alert the practitioner to the possibility that treatment effects vary widely from subject to subject in the population but also quantify the risk involved by providing suitable confidence bounds. A final decision concerning the application of the treatment to a subject must of course be based on the results of statistical analyses together with any subject matter knowledge that may be available.

APPENDIX

Derivation of Eq. (4) is as follows. Since X and Y are only observable for different subjects, any estimator computed from observed X will be independent from one computed from observed Y. So $\hat{\sigma}^2_X$ and $\hat{\sigma}^2_Y$ are independent and asymptotically normal with mean σ^2_X and σ^2_Y, respectively, and variance matrix

$$\boldsymbol{V} = \begin{pmatrix} \frac{2}{n_1}\sigma^4_X & 0 \\ 0 & \frac{2}{n_2}\sigma^4_Y \end{pmatrix}$$

Define $J = (J_1, J_2)$ where

$$J_1 = \frac{\partial\sigma^2_D}{\partial\sigma^2_X} = 1 - \rho_{XY}\sigma_Y/\sigma_X$$

$$J_2 = \frac{\partial\sigma^2_D}{\partial\sigma^2_Y} = 1 - \rho_{XY}\sigma_X/\sigma_Y$$

Then, for a fixed ρ_{XY}, the asymptotic distribution of $\hat{\sigma}^2_D$ is normal with mean σ^2_D and variance computed by the matrix multiplication

$$\mathrm{Var}(\hat{\sigma}^2_D) = \boldsymbol{J}\ \boldsymbol{V}\ \boldsymbol{J}^T$$

Derivation of Eq. (5) proceeds in a similar manner. The joint distribution of $(\hat{\mu}_D, \hat{\sigma}^2_D)^T$ is asymptotically normal with mean vector $(\mu_D, \sigma^2_D)^T$ and variance matrix,

$$\boldsymbol{U} = \begin{pmatrix} \mathrm{Var}(\hat{\mu}_D) & 0 \\ 0 & \mathrm{Var}(\hat{\sigma}^2_D) \end{pmatrix}$$

where $\mathrm{Var}(\hat{\mu}_D) = \sigma^2_X/n_1 + \sigma^2_Y/n_2$ and $\mathrm{Var}(\hat{\sigma}^2_D)$ is given in Eq. (4). Recall $P_- = \Phi(-\mu_D/\sigma_D)$, and define $M = (M_1, M_2)$ where

$$M_1 = \frac{\partial P_-}{\partial\mu_D} = \frac{-\phi(\mu_D/\sigma_D)}{\sigma_D}$$

$$M_2 = \frac{\partial P_-}{\partial\sigma^2_D} = \frac{-\phi(\mu_D/\sigma_D)\mu_D}{2\sigma^3_D}$$

Then the distribution of $\hat{P}_-$ is asymptotically normal with mean P_- and variance computed by $M\ \boldsymbol{U}\ M^T$.

ACKNOWLEDGMENTS

The authors thank the editor, an associate editor, and two anonymous reviewers for helpful comments.

REFERENCES

1. Cox, D.R. The interpretation of the effects of non-additivity in the Latin square. Biometrika **1958**, *45*, 69–73.
2. Longford, N.T. Selection bias and treatment heterogeneity in clinical trials. Statistics in Medicine. **1999**, *18*, 1467–1474.
3. Rioux, P.P. Clinical trials in pharmacogenetics and pharmacogenomics: methods and applications. American Journal of Health-System Pharmacy **2000**, *57*, 887–898.
4. Allison, D.B.; Mentore, J.M.; Heo, M.; Chandler, L.; Cappelleri, L.C.; Infante, M.; Weiden, P. Meta-analysis of the effects of anti-psychotic medication on weight gain. American Journal of Psychiatry **1999**, *156*, 1686–1696.
5. Rietschel, M.; Naber, D.; Fimmers, R.; Moller, H.J.; Propping, P.; Nothen, M.M. Efficacy and side-effects of clozapine not associated with variation in the 5-HT2C receptor. Neuroreport **1997**, *8*, 1999–2003.
6. Holland, P.W. Statistics and causal inference (with discussion). Journal of the American Statistical Association **1986**, *81*, 945–970.
7. Gadbury, G.L.; Iyer, H.K. Unit-treatment interaction and its practical consequences. Biometrics **2000**, *56*, 882–885.
8. Rubin, D.B. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology **1974**, *66*, 688–701.
9. Cox, D.R. *Planning of Experiments*; John Wiley & Sons: New York, 1958.
10. Tukey, J.W. One degree of freedom for nonadditivity. Biometrics **1949**, *5*, 232–242.
11. Wilk, M.B.; Kempthorne, O. Non-additivities in a Latin Square Design. Journal of the American Statistical Association **1957**, *52*, 218–236.
12. Hinkelmann, K.; Kempthorne, O. *Design and Analysis of Experiments*, Vol. 1; John Wiley & Sons: New York, 1994.
13. Hollander, M.; Wolfe, D.A. *Nonparametric Statistical Methods*; John Wiley & Sons: New York, 1999.
14. Lord, F.M. Estimation of parameters from incomplete data. Journal of the American Statistical Association **1955**, *50*, 870–876.
15. Moore, D.S.; McCabe, G.P. *Introduction to the Practice of Statistics*, 3rd Ed.; W.H. Freeman and Company: New York, 1999.
16. Efron, B.; Tibshirani, R.J. *An Introduction to the Bootstrap*; Chapman & Hall: New York, 1993.
17. Gadbury, G.L. Randomization inference and bias of standard errors. The American Statistician **2001**, *55*, 310–313.

Received May 2001
Revised September, November 2001
Accepted November 2001