Biometrical Journal **47** (2005) 5, 662–673 DOI: 10.1002/bimj.200510157

Assessing Treatment Effect Heterogeneity in Clinical Trials with Blocked Binary Outcomes

Jeffrey M. Albert[*, 1], **Gary L. Gadbury**[2], and **Edward J. Mascha**[3]

[1] Department of Epidemiology and Biostatistics, Case Western Reserve University, 10900 Euclid Ave., Cleveland, OH, 44106, USA

[2] Department of Mathematics and Statistics, University of Missouri – Rolla, Rolla, MO, 65409, USA

[3] Collaborative Biostatistics Center, Cleveland Clinic Foundation 9500 Euclid Ave., Cleveland, OH, 44195, USA

Received 3 March 2005, revised 22 April 2005, accepted 6 May 2005

Summary

This paper addresses treatment effect heterogeneity (also referred to, more compactly, as 'treatment heterogeneity') in the context of a controlled clinical trial with binary endpoints. Treatment heterogeneity, variation in the true (causal) individual treatment effects, is explored using the concept of the potential outcome. This framework supposes the existance of latent responses for each subject corresponding to each possible treatment. In the context of a binary endpoint, treatment heterogeniety may be represented by the parameter, π_2, the probability that an individual would have a failure on the experimental treatment, if received, and would have a success on control, if received. Previous research derived bounds for π_2 based on matched pairs data. The present research extends this method to the blocked data context. Estimates (and their variances) and confidence intervals for the bounds are derived. We apply the new method to data from a renal disease clinical trial. In this example, bounds based on the blocked data are narrower than the corresponding bounds based only on the marginal success proportions. Some remaining challenges (including the possibility of further reducing bound widths) are discussed.

Key words: Bounds; Causal effects; Counterfactuals; Potential outcomes; Randomized block design; Subject-treatment interaction.

1 Introduction

In a standard phase III clinical trial, patients are randomized to either a new treatment or a standard (or control) treatment, and followed for relevant health endpoints. Ideally, data from such a study will allow future patients and their doctors to make an informed choice between the competing therapies. Statistical reports from clinical trials typically present the difference (or ratio) of means (or proportions) between the two study groups. The implicit assumption is that this estimated population effect represents a common treatment effect for each patient. Thus most reports and their readers overlook the possibility – one of both clinical and scientific import – that patients may have heterogeneous treatment effects.

To clarify issues concerning individual treatment effects, we use the concept of the potential outcome (Rubin, 1974; Rosenbaum and Rubin, 1983a; Holland, 1986). In a clinical trial comparing two interventions (experimental treatment versus control, say), we may conceive of two possible responses for an individual – namely a response if the individual were to receive the experimental treatment and a response if the individual were to receive the control intervention. Both these re-

[*] Corresponding author: e-mail: jma13@case.edu

sponses are viewed as existing, albeit latently, and are referred to as potential outcomes. Of course, only one of these outcomes is eventually observed (corresponding to the treatment assigned to that individual), while the other response remains unknown (and is thus sometimes referred to as a 'counterfactual').

We let $Y_h(u)$ denote the response (potential outcome) for patient u if given treatment h, where, in the present context, $h = T$ (treatment) or C (control). This subject then has (causal) treatment effect $\alpha(u) = Y_T(u) - Y_C(u)$. In the standard randomized comparison trial we observe either $Y_T(u)$ or $Y_C(u)$ but not both. Thus $\alpha(u)$ is unknown, and may be viewed as a realization of the latent random variable, α. Although the crossover trial design (as employed in bioequivalence studies – see, for example, Sheiner, 1992 and Schall, 1995) has been invoked as a solution to this problem, this design is not feasible in many cases – particularly, when binary responses (such as 'cure'/'no cure') are of interest. Furthermore, the responses for an individual are not observed at the same time and thus additional assumptions are required to make inferences regarding causal treatment effects.

Using this potential outcomes framework, we may write the mean treatment effect, the focus of most clinical trial analyses, as $E(\alpha)$. We interpret expectation in this context as the average over the population; note that we drop the individual argument to indicate the corresponding random variable. Variation in treatment effects may be quantified by $V(\alpha)$ (the variance of α) or, for instance, by $P(\alpha < 0)$, the proportion of individuals with a worse outcome on treatment than control (assuming larger values of Y indicate a better response). Longford (1999) labelled the occurrence of non-zero variability in treatment effects (that is, $V(\alpha) > 0$) as 'treatment heterogeneity'. It has also been referred to as 'unit-treatment interaction' (Gadbury and Iyer, 2000) or 'subject-treatment interaction' (Gadbury, Iyer, and Allison, 2001).

In this paper we address the case, common in clinical trials, where the response, Y, is a univariate binary variable, with $Y = 1$ indicating success and $Y = 0$ indicating failure. Here, the possible values of the treatment effect for an individual (α) are -1, 0, and 1. The limitation of an analysis based solely on the difference in success proportions is easily illustrated. Suppose that population success rates for the new and standard treatments are 0.7 and 0.5, respectively, so that the mean treatment effect is $E(\alpha) = E(Y_T - Y_C) = 0.2$. This result does not reveal the degree of heterogeneity of individual treatment effects. For instance, one possible situation is that $P(\alpha = 1) = 0.2$ and $P(\alpha = 0) = 0.8$; that is, 20% of the patients would benefit from the new treatment (relative to the standard), while treatment would have no effect for the remaining 80%, (that is, these patients would have the same response on either the new or standard treatment). Alternatively, it may be that $P(\alpha = 1) = 0.5$, $P(\alpha = 0) = 0.2$, and $P(\alpha = -1) = 0.3$, so that 50% of the patients would benefit from new treatment (relative to the standard) but 30% would have a worse outcome on new relative to standard treatment (and the remaining 20% would have zero effect). These two scenarios, representing minimum and maximum treatment effect heterogeneity, respectively, would have different implications for the clinical use of the new (relative to the standard) intervention.

One approach for resolving this indeterminacy is to make assumptions regarding the probabilities of these causal effects. For example, a common assumption in causal inference is that of 'monotone treatment effect', variants of which were considered, for example, by Goetghebeur and Molenberghs (1996) and Angrist, Imbens, and Rubin (1996). In the present context, monotonicity implies that the potential outcome of an individual on the new treatment would be no worse than the potential outcome of the individual on control. Making this assumption (and thereby 'assuming away the problem') would defeat the purpose of an investigation of treatment heterogeneity. Furthermore, in many applications there are scientific and empirical reasons to suspect the existence of a substantial proportion of patients who would respond on control but fail to respond on active treatment. In their discussion of run-in trials, Berger, Rezvani and Makarewicz (2003) suggested some reasons for the existance of individuals who would respond even to an 'inactive' control but not to the experimental treatment. They noted, in particular, that when the interventions involve ancillary treatment or attention: 1) an individual may receive superior ancillary care on control than on the experimental treatment, or 2) an individual who suffers side effects from the experimental treatment may dropout or otherwise lose the benefit of the ancillary treatment. Berger et al. also pointed out that placebo has been shown to be superior to active treatments in some trials (for example, the CAST Trial, 1989; Prystowsky, Katz, and Knilans, 1990). As one of their key findings, Berger et al. demonstrated the important implications of the subject-treatment interaction (in particular, the occurrence or non-occurrence of subjects who would respond to control but not to experimental treatment) in evaluating the effect of run-in trials.

Methods for assessing subject-treatment interaction have only recently begun to appear in the literature. Gadbury and Iyer (2000) provided bounds for $V(\alpha)$ in the case of a normally-distributed continuous endpoint by making use of covariate information. Gadbury et al. (2001) further studied this situation, focusing on a sensitivity analysis as a practical approach to assessing the range of possible treatment heterogeneity. Gadbury, Iyer, and Albert (2004) addressed the case of a binary response in the context of a matched pairs design. This research, discussed further below, provided a bound for $P(\alpha < 0)$, the tightness of which depends on the quality of the matching.

Note that when effect modifiers (or variables that explain differential treatment effects) are available, established methods to assess covariate-treatment interactions may be utilized. The subject-treatment interaction, in contrast, refers to *unexplained* treatment effect heterogeneity, the assessment of which may be of interest in the absence of, or conditional on, known effect modifiers.

The current paper extends the work of Gadbury et al. (2004) (abbreviated as GIA hereon in) to the more general setting of blocked data. Following the background discussion in Section 2, we present in Section 3 bounds for $P(\alpha < 0)$ that make use of blocked data, and propose estimators for these bounds. In Section 4, we provide estimators of the variances for the bound estimators, as well as confidence intervals for the bounds. Section 5 illustrates the method using data from a renal disease clinical trial. Section 6 provides concluding comments.

2 Background

In the present context of a two-arm randomized clinical trial, there are two potential outcome variables: Y_T, Y_C. These outcomes are defined for each subject (even if not observed) and may be considered as inherent, baseline variables. Since our outcomes are binary, we may thus conceive of each subject as falling into one of four populations ($\Pi_1, \dots, \Pi_4$, say), whose corresponding outcomes and probabilities (or population proportions) are given in the following table:

(y_T, y_C)	(0,0)	(0,1)	(1,0)	(1,1)
$P(Y_T = y_T, Y_C = y_C)$	π_1	π_2	π_3	π_4

. (1)

We focus on the unknown parameter π_2, the proportion of patients who would have a worse outcome on treatment than on control (expressed above as $P(\alpha < 0)$). A similar development is possible for the parameter π_3 (the probability that treatment is more beneficial than control for an individual). Note that although certain linear combinations of the π's are estimable, none of the four probabilities are individually identifiable. We denote the marginal (and estimable) success probabilities as $p_h = P(Y_h = 1)$ for $h = T, C$. GIA give general bounds for π_2 (which follow from the constraints of the 2 by 2 contingency table; that is, $\pi_1 + \pi_2 = p_C$ and $\pi_1 + \pi_3 = p_T$):

$$\max\,(0, p_C - p_T) \equiv L \le \pi_2 \le U \equiv \min\,(1 - p_T, p_C). \tag{2}$$

Estimates of the bounds are obtained by substitution of corresponding sample proportions (thus providing maximum likelihood estimates). To illustrate, we suppose as in the example discussed above, that $\hat{p}_T = 0.7$ and $\hat{p}_C = 0.5$. From (2), we obtain estimated bounds for π_2 as: $\hat{L} = 0$ and $\hat{U} = 0.3$.

The 'simple' bounds for π_2, given by (2), will typically be rather conservative. GIA sought tighter bounds by considering the context of a matched pairs design. In the standard version of the design, one member of each pair is randomly selected (each member with probability 0.5) to receive (new) treatment whereby the other member receives control. GIA also considered an extended matched pairs design in which the members of some pairs both receive treatment and other pairs both receive control (although bounds may be constructed if only double treatment or only double control pairs are considered). For any matched subjects u_1 and u_2, GIA defined the following (estimable) probabilities:

$$\begin{aligned} g_2 &= P(Y_T(u_1) = 0, Y_C(u_2) = 1), \\ h_T &= P(Y_T(u_1) = 1, Y_T(u_2) = 1), \\ h_C &= P(Y_C(u_1) = 1, Y_C(u_2) = 1) \end{aligned} \tag{3}$$

Thus, g_2 represents the probability that in a randomized matched pair, the subject assigned to treatment fails and the subject assigned to control succeeds; $h_T(h_C)$ is the probability that in a matched pair with both members assigned to treatment (control) both would have a success. A refined set of bounds for π_2 is obtained (Gadbury et al., 2004) as:

$$\begin{aligned} L_M &\equiv \max\,(0, g_2 - \min\,(p_T - h_T, p_C - h_C)) \\ U_M &\equiv g_2 + \min\,(p_T - h_T, p_C - h_C). \end{aligned} \tag{4}$$

If matching occurs at random (that is, $h_T = p_T{}^2$, $h_C = p_C{}^2$, and $g_2 = (1 - p_T)\, p_C$) then $U_M \geq U$; thus, there would be no benefit to matching. However, 'better-than-random' matching has the potential to tighten the bounds for π_2 relative to the simple bounds (2). Note that while g_2 is estimable in the standard matched pairs design, the extended matched pairs design is required to estimate h_T and/or h_C and thus take practical advantage of the refined theoretical bounds.

3 Bounds for Blocked Data

3.1 Conceptualization and derivation of bounds

Since the matched pairs design is not commonly used in clinical trials in practice, it would be useful to extend the results to more general circumstances. The development in this section is based on the idea that bounds may be constructed making use of blocks (each containing subjects assigned to both new and standard treatments). A matched pair may be viewed as a special case of a block. We may, for example, make use of randomized block designs (Neter, Wasserman, and Kutner, 1985) in which treatments are randomized within natural blocks such as families. A more common design involves the use of permuted blocks (where a block comprises subjects entering at proximal times), often within strata. In addition blocks may be obtained from what would technically be considered stratified randomization (where the block size is not pre-determined); examples of such blocks include clinical center or physician. Groups based on other stratifying variables such as age and disease status do not typically exhibit the exchangeability property often attributed (in statistical models) to blocks. However, such groups may be used as blocks in our method. Even broader applicability of our approach is obtained by recognizing that it may utilize blocks constructed post-randomization.

In our extension, instead of considering subjects u_1 and u_2 as a matched pair, we consider them as belonging to the same block. With this new interpretation, the discordance and concordance probabilities, g_2, h_T, and h_C are defined as in (3). For instance, g_2 represents the probability that, for a pair of subjects in the same block, one on treatment and the other on control, the subject on treatment will experience a failure and the subject on control will experience a success ('control beating treatment'). This definition and corresponding estimate assume that treatment assignment is unrelated to the potential outcomes. This condition, referred to as strong ignorability of treatment assignment (or SITA, Rosenbaum and Rubin, 1983b) follows from *effective* randomization of treatments. Of course, SITA may be violated with departures from the intended randomization mechanism (for example, as a result of selection bias).

Our primary interest is in obtaining bounds for π_2, the probability that an individual's potential outcome on treatment would be worse than that on control. Our population model is again given by (1).

We assume, analogously to the matched-pairs analysis, that subjects are exchangeable within blocks and that within-block probabilities are constant across blocks. As in Gadbury et al. (2004), we assume Rubin's stable unit treatment value assumption (SUTVA, Rubin, 1980) – that the potential outcomes of an individual are not affected by the treatment received by others.

The bounds for the blocked data are derived in an analogous manner to those for matched pairs data and take a similar form (but where parameters have the interpretation relevant to the blocked context). We write the bounds as:

$$\begin{aligned} L_B &\equiv \max\left(0, g_2 - \min\left(p_T - h_T, p_C - h_C\right)\right), \\ U_B &\equiv g_2 + \min\left(p_T - h_T, p_C - h_C\right). \end{aligned} \tag{5}$$

A simple derivation of these bounds is provided in the Appendix.

3.2 Estimation of bounds

Estimation in the blocked case is not as straight-forward as for matched pairs. We estimate probabilities (occurring in (5)) using a nonparametric approach that involves the counting of all possible pairs (within blocks). First, some additional notation is needed. Let n_{hij} be the number of subjects in block j ($j = 1, \dots, m$) on treatment h who have outcome i ($i \in \{0,1\}$). Also, let n_{hj} be the total number of subjects in block j on treatment h. Estimators of probabilities are obtained as the ratio of the number of pairs of subjects (in the same block) with the indicated outcomes divided by the total number of pairs. Thus our estimators of the probability of 'control beating treatment' (g_2) and the probabilities of concordance for a pair on the same treatment (h_T, h_C) are,

$$\hat{g}_2 = \frac{\sum\limits_j n_{C1j} n_{T0j}}{\sum\limits_j n_{Cj} n_{Tj}},$$

$$\hat{h}_T = \frac{\sum\limits_j n_{T1j}(n_{T1j} - 1)}{\sum\limits_j n_{Tj}(n_{Tj} - 1)},$$

$$\hat{h}_C = \frac{\sum\limits_j n_{C1j}(n_{C1j} - 1)}{\sum\limits_j n_{Cj}(n_{Cj} - 1)}.$$

Estimates of p_T and p_C are easily obtained as the (marginal) sample proportions of successes on treatment and control, and are denoted as $\hat{p}_T = \frac{\sum\limits_j n_{T1j}}{\sum\limits_j n_{Tj}}$ and $\hat{p}_C = \frac{\sum\limits_j n_{C1j}}{\sum\limits_j n_{Cj}}$, respectively. Substitution of estimates for unknown parameters in (5) yields estimated bounds, $\hat{L}_B$ and $\hat{U}_B$.

4 Variance Estimators for Bounds

To make practical use of the estimated bounds, it is important to have an idea of their sampling variability. In this section we provide approximate confidence intervals based on estimated variances. To begin, we consider the variance of $\hat{U}_B$ (denoted $V(\hat{U}_B)$). From (5), after substitution of estimates, we can write $\hat{U}_B = I_m(\hat{g}_2 + (\hat{p}_T - \hat{h}_T)) + (1 - I_m)\,(\hat{g}_2 + (\hat{p}_C - \hat{h}_C))$ where $I_m = I[\hat{p}_T - \hat{h}_T \le \hat{p}_C - \hat{h}_C]$; that is, I_m is an indicator variable that takes value 1 if the condition in the brackets is met and takes value 0 otherwise. As an approximation, and to provide a feasible formula for the variance of $\hat{U}_B$, we take I_m to be known (that is, for the condition to have probability 0 or 1 given the marginal probabilities). Thus, we use as an approximate formula for $V(\hat{U}_B)$,

$$V_{U_B} \equiv I_m V_{TU} + (1 - I_m)\, V_{CU} \tag{6}$$

where $V_{TU} = V(\hat{g}_2 + (\hat{p}_T - \hat{h}_T))$ and $V_{CU} = V(\hat{g}_2 + (\hat{p}_C - \hat{h}_C))$. We expect V_{U_B} to be a good approximation to $V(\hat{U}_B)$, since the cases in which $P(I_m = 1)$ departs most from 0 or 1 (and gets close to 0.5, say) are ones in which the values of $\hat{p}_T - \hat{h}_T$ and $\hat{p}_C - \hat{h}_C$ are similar and thus the value of the estimate $(\hat{U}_B)$ not greatly affected by the value of I_m.

An exact expression for V_{TU} is obtained as follows:

$$V_{TU} = V(\hat{g}_2) + V(\hat{p}_T) + V(\hat{h}_T) + 2[C(\hat{g}_2, \hat{p}_T) - C(\hat{g}_2, \hat{h}_T) - C(\hat{p}_T, \hat{h}_T)] \tag{7}$$

where

$$\begin{aligned}
V(\hat{h}_T) &= (n_{T2\bullet}^{-2})\, \{h_T(1 - h_T)\, n_{T2\bullet} + 2[(h_{T2} - h_{TT})\, n_{T22\bullet} + (h_{TT} - h_T^2)\, n_{T4\bullet}]\},\\
V(\hat{g}_2) &= (n_{TC\bullet}^{-2})\, \{g_2(1 - g_2)\, n_{TC\bullet} + 2[(g_{2\bar{T}} - g_2^2)\, n_{T2C\bullet} + (g_{2C} - g_2^2)\, n_{TC2\bullet}]\\
&\qquad + 2(g_{2\bar{T}C} - g_2^2)\, n_{T2C2\bullet}\},\\
V(\hat{p}_T) &= (n_T^{-2})\, \{p_T(1 - p_T)\, n_T + 2(h_T - p_T^2)\, n_{T2\bullet}\},\\
C(\hat{g}_2, \hat{p}_T) &= (n_T n_{TC\bullet})^{-1}\, \{2(g_{2T} - p_T g_2)\, n_{T2C\bullet} - p_T g_2 n_{TC\bullet}\},\\
C(\hat{g}_2, \hat{h}_T) &= (n_{T2\bullet} n_{TC\bullet})^{-1}\, \{(g_{2TT} - h_T g_2)\, n_{T2TC\bullet} - 2 g_{2TT} n_{T2C\bullet}\},\\
C(\hat{p}_T, \hat{h}_T) &= (n_T n_{T2\bullet})^{-1}\, \{(h_{TT} - p_T h_T)\, n_{T2T\bullet} - 2(h_{TT} - h_T)\, n_{T2\bullet}\},\\
n_{T2j} &= \frac{n_{Tj}(n_{Tj} - 1)}{2}, \qquad n_{C2j} = \frac{n_{Cj}(n_{Cj} - 1)}{2}, \qquad n_{T4j} = \frac{n_{T2j}(n_{T2j} - 1)}{2},\\
n_{T22j} &= \frac{n_{Tj}(n_{Tj} - 1)\,(n_{Tj} - 2)\,(n_{Tj} - 3)}{8}, \qquad n_{T2TCj} = n_{T2j} n_{Tj} n_{Cj},\\
n_{TCj} &= n_{Tj} n_{Cj}, \qquad n_{TC2j} = n_{Tj} n_{C2j}, \qquad n_{T2Cj} = n_{T2j} n_{Cj},\\
n_{T2Tj} &= n_{T2j} n_{Tj}, \qquad n_{C2Cj} = n_{C2j} n_{Cj}, \qquad n_{T2C2j} = n_{T2j} n_{C2j},
\end{aligned}$$

the use of a dot in place of j indicates the sum over j (blocks), and for different (exchangeable) subjects $(u_1, \ldots, u_4)$ in the same block,

$$\begin{aligned}
g_{2T} &= P(Y_T(u_1) = 0, \quad Y_C(u_2) = 1, \quad Y_T(u_3) = 1),\\
g_{2TT} &= P(Y_T(u_1) = 0, \quad Y_C(u_2) = 1, \quad Y_T(u_3) = 1, \quad Y_T(u_4) = 1),\\
g_{2\bar{T}} &= P(Y_T(u_1) = 0, \quad Y_C(u_2) = 1, \quad Y_T(u_3) = 0),\\
g_{2C} &= P(Y_T(u_1) = 0, \quad Y_C(u_2) = 1, \quad Y_C(u_3) = 1),\\
g_{2\bar{T}C} &= P(Y_T(u_1) = 0, \quad Y_C(u_2) = 1, \quad Y_T(u_3) = 0, \quad Y_C(u_4) = 1),\\
h_{TT} &= P(Y_T(u_1) = 1, \quad Y_T(u_2) = 1, \quad Y_T(u_3) = 1),\\
h_{T2} &= P(Y_T(u_1) = 1, \quad Y_T(u_2) = 1, \quad Y_T(u_3) = 1, \quad Y_T(u_4) = 1)\,.
\end{aligned}$$

A sketch of the derivation of $V(\hat{h}_T)$, one of the terms in (7), is given in the Appendix. Estimates for other terms are similarly obtained. The expression for V_{TU} (7) contains unknown parameters (within-block joint probabilities). We obtain an estimator, $\hat{V}_{TU}$, by substitution of estimates for the unknown probabilities. The estimators for some of these, namely, g_2, h_T, and h_C, have already been given (Section 3.2). Other estimators are obtained in a similar manner (extended to deal with joint probabilities involving more than two subjects). For instance, probabilities involving three subjects (in a block) are estimating by counting the number of possible triplets with the indicated outcomes (and dividing by the total number of triplets on the indicated treatments). For reference, these estimators are given as follows:

$$\hat{g}_{2T} = \frac{2 \sum\limits_j n_{T0j} n_{C1j} n_{T1j}}{\sum\limits_j n_{Tj}(n_{Tj}-1)\, n_{Cj}}, \qquad \hat{g}_{2TT} = \frac{3 \sum\limits_j n_{T0j} n_{C1j} n_{T1j}\, (n_{T1j}-1)}{\sum\limits_j n_{Tj}(n_{Tj}-1)\,(n_{Tj}-2)\, n_{Cj}},$$

$$\hat{g}_{2\bar{T}} = \frac{\sum\limits_j n_{T0j}(n_{T0j}-1)\, n_{C1j}}{\sum\limits_j n_{Tj}(n_{Tj}-1)\, n_{Cj}}, \qquad \hat{g}_{2C} = \frac{\sum\limits_j n_{T0j} n_{C1j}(n_{C1j}-1)}{\sum\limits_j n_{Tj} n_{Cj}(n_{Cj}-1)},$$

$$\hat{g}_{2\bar{T}C} = \frac{\sum\limits_j n_{T0j}(n_{T0j}-1)\, n_{C1j}(n_{C1j}-1)}{\sum\limits_j n_{Tj}(n_{Tj}-1)\, n_{Cj}(n_{Cj}-1)},$$

$$\hat{h}_{TT} = \frac{\sum\limits_j n_{T1j}(n_{T1j}-1)\,(n_{T1j}-2)}{\sum\limits_j n_{Tj}(n_{Tj}-1)\,(n_{Tj}-2)}, \qquad \hat{h}_{T2} = \frac{\sum\limits_j n_{T1j}(n_{T1j}-1)\,(n_{T1j}-2)\,(n_{T1j}-3)}{\sum\limits_j n_{Tj}(n_{Tj}-1)\,(n_{Tj}-2)\,(n_{Tj}-3)}. \tag{8}$$

An expression for V_{CU} (also involved in the right side of (6)) is obtained in a similar manner to V_{TU}.

Thus, we have,

$$V_{CU} = V(\hat{g}_2) + V(\hat{p}_C) + V(\hat{h}_C) + 2[C(\hat{g}_2, \hat{p}_C) - C(\hat{g}_2, \hat{h}_C) - C(\hat{p}_C, \hat{h}_C)] \tag{9}$$

where $V(\hat{p}_C)$, $V(\hat{h}_C)$, and $C(\hat{p}_C, \hat{h}_C)$ are obtained with "*C*" substituted for "*T*" in the corresponding formulae and definitions given above; and

$$C(\hat{g}_2, \hat{p}_C) = (n_C n_{TC\bullet})^{-1} \{2(g_{2C} - p_C g_2)\, n_{TC2\bullet} + g_2(1-p_C)\, n_{TC\bullet}\},$$

$$C(\hat{g}_2, \hat{h}_C) = (n_{C2\bullet} n_{TC\bullet})^{-1} \{(g_{2CC} - h_C g_2)\, n_{C2TC\bullet} + 2(g_{2C} - g_{2CC})\, n_{TC2\bullet}\}$$

where

$$n_{C2TCj} = n_{C2j} n_{Tj} n_{Cj}, \quad g_{2CC} = P(Y_T(u_1) = 0, \quad Y_C(u_2) = 1, \quad Y_C(u_3) = 1, \quad Y_C(u_4) = 1),$$

for different (exchangeable) subjects in the same block, $u_1, \ldots, u_4$. We obtain the estimator $\hat{V}_{CU}$ after substitution of estimates as given above, as well as the estimate for g_{2CC},

$$\hat{g}_{2CC} = \frac{\sum\limits_j n_{C1j}(n_{C1j}-1)\,(n_{C1j}-2)\, n_{T0j}}{\sum\limits_j n_{Cj}(n_{Cj}-1)\,(n_{Cj}-2)\, n_{Tj}}$$

for the corresponding parameters in V_{CU} (9).

Note that a given block must have at least four subjects on each treatment in order to contribute to the estimates of all the terms in the formula for V_{U_B}(6). This is seen, specifically, in the expression for $\hat{h}_{T2}$ (8) (similarly, $\hat{h}_{C2}$). Other terms require fewer subjects (two or three, possibly treatment-specific) from each block.

Using (6), and substituting estimates given above, we obtain our estimate for $V(\hat{U}_B)$ as:

$$\hat{V}(\hat{U}_B) = I_m \hat{V}_{TU} + (1 - I_m)\, \hat{V}_{CU}. \tag{10}$$

The estimate of the variance for the lower bound is readily obtained from the above results. In particular, we obtain, using the approximate formula for $V(\hat{L}_B)$ analgous to (6), the estimate

$$\hat{V}(\hat{L}_B) = I_m \hat{V}_{TL} + (1 - I_m)\, \hat{V}_{CL} \tag{11}$$

where $\hat{V}_{TL}$, $\hat{V}_{CL}$ are obtained by substituting estimates (using formulae given above) for unknown parameters in the corresponding variance expressions, $V_{TL} = V(\hat{g}_2 - (\hat{p}_T - \hat{h}_T))$, and $V_{CL} = V(\hat{g}_2 - (\hat{p}_C - \hat{h}_C))$.

Approximate confidence intervals may be obtained based on the above variances (and using a normal approximation for the distribution of the bound estimators). For instance, a $100(1 - \alpha)\%$ confidence interval for the upper bound, U_B, is obtained as

$$\hat{U}_B \pm z_{1-\alpha/2}(\hat{V}(\hat{U}_B))^{1/2}$$

where $z_{1-\alpha/2}$ is the $(100(1 - \alpha/2)$ percentile of the standard normal distribution. A confidence interval for the lower bound L_B, is obtained similarly. Note that the confidence intervals for the bounds provide a conservative confidence interval for π_2.

5 Data Application

We apply the preceding method to data from the Modified Diet in Renal Disease (MDRD) Study (MDRD Study Group, 1995), a multicenter, 2 × 2 factorial randomized study of the effects of diet and blood pressure interventions on long-term kidney function in chronic renal disease patients. A primary

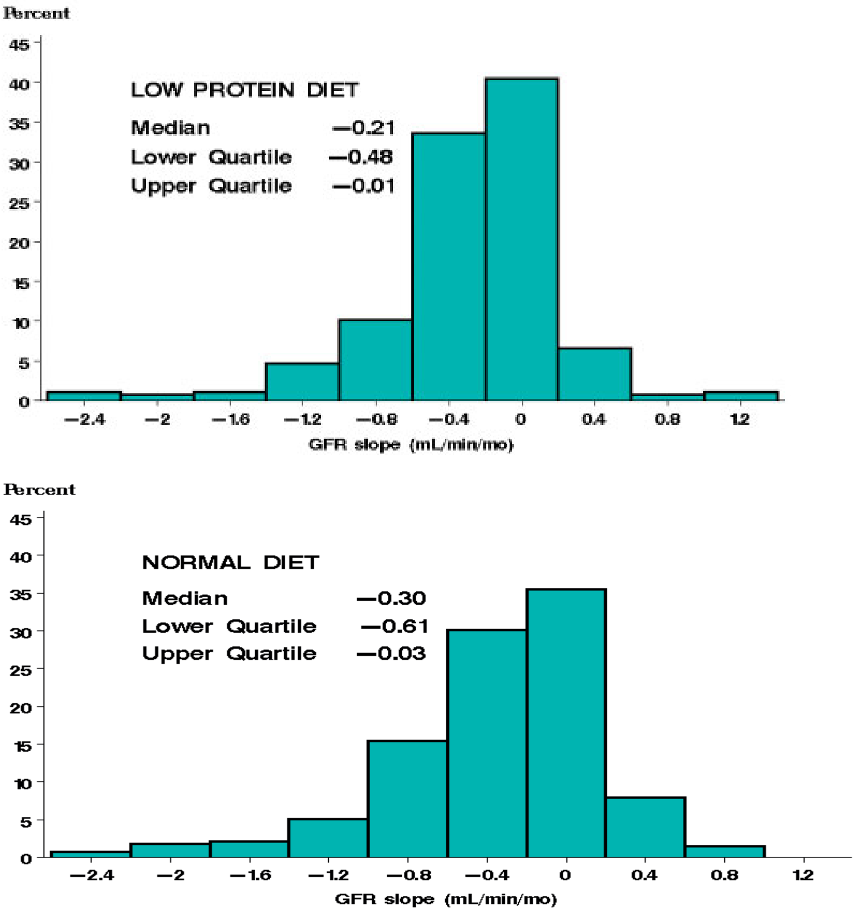

Figure 1 Histograms of GFR slope (ml/min/mo) for each treatment group.

end point was the slope of the glomerular filtration rate (GFR) over time, starting four months after randomization. We focus here on the diet component of the study which involved the comparison of a low protein to a normal protein diet. It was hypothesized that a low protein diet (relative to a normal protein diet) would delay the progression of disease – thus producing a greater mean slope for GFR. To create a binary outcome, we used the median value for the GFR slope as a cutpoint; this was suggested to us as a reasonable cutpoint by one of the MDRD principal investigators (Tom Greene, Ph.D., personal correspondence). A GFR slope greater than the median value of -0.23 was thus considered a 'success', while a lower slope was labelled a 'failure'.

Histograms of GFR slopes for each treatment group are presented in Figure 1. The estimated rate of success (GFR slope greater than -0.23) on the low protein diet was $\hat{p}_T = 0.53$ ($n_T = 261$); for the normal protein diet, the success rate was $\hat{p}_C = 0.47$ ($n_C = 265$). A test for a difference in success rates between the two groups, based on GEE (with an exchangeable correlation structure for subjects within center, nested within treatment group) was not significant ($P = 0.15$). Although the low protein diet did not show a benefit in terms of proportion with GFR slope above the median, it was of interest to investigate treatment effect heterogeneity across individuals.

A preliminary step to assessing treatment heterogeneity in our method is to define blocks. Although clinical center (of which there were fifteen in the MDRD study) provided a natural block, the within-center correlation was rather low (0.009). We thus constructed, post-randomization, blocks using baseline prognostic variables, namely, GFR CV (coefficient of variation based on four baseline measurements), polycystics (yes/no), urine protein, urine creatinine, serum urea nitrogen, total cholesterol, iron, hemoglobin A1c, reported protein and energy intakes, potassium, magnesium, white blood count, mean arterial pressure, systolic blood pressure, unadjusted and adjusted creatinine clearance, and weight. Logistic regression was used to obtain a score representing the predicted logit of success as a function of the 18 covariates (ignoring treatment assignment). From this score, using appropriate percentiles, forty equal sized blocks were constructed.

An empirical description of treatment effect heterogeneity is provided in Figure 2 which plots estimated treatment effect (the difference in success proportions for treatment versus control) by block (ordered by the predicted probability of success). The plot shows the substantial proportion of blocks ($13/40 = 0.325$) with higher success rates on control (normal protein diet) than on the experimental treatment (low protein diet), that is, below the line at $\hat{p}_T - \hat{p}_C = 0$. To assess treatment effect heterogeneity more rigorously, we employed the method described above to estimate bounds on the probability that an individual would succeed on the normal protein, but fail on the low protein diet treatment (π_2).

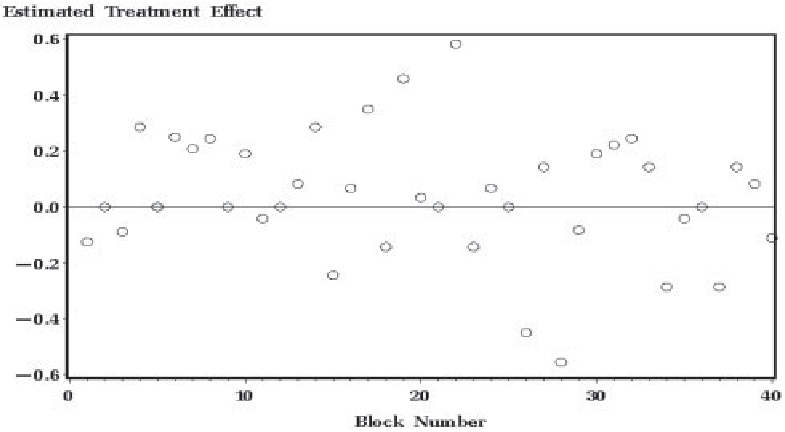

Figure 2 Estimated treatment effect ($\hat{p}_T - \hat{p}_C$) by block number (blocks in order of predicted probability of success).

The estimated simple bounds for π_2 (from (2)) are (0, 0.47). The blocked data (using (5)) produced estimated bounds for π_2 of (0, 0.41). The bounds are thus somewhat narrower in this example when we make use of blocked relative to the unblocked data. The variance for the lower and upper (blocked data) bounds are calculated from (10) and (11) as $\hat{V}(\hat{L}_B) = 0.00028$ and $\hat{V}(\hat{U}_B) = 0.00128$, respectively. A conservative 95% confidence interval for π_2 may thus be obtained as $(\min(0, \hat{L}_B - 1.96[\hat{V}(\hat{L}_B)]^{1/2}), \max(1, \hat{U}_B + 1.96[V(\hat{U}_B)]^{1/2})) = (0,\ 0.46)$. By way of comparison, we would find under independence of potential outcomes (under treatment or control) an estimated probability of a treatment failure and control success (π_2) of $(1 - \hat{p}_T)\,\hat{p}_C = (0.47)\,(0.47) = 0.22$. The value $\pi_2 > 0.22$ thus represents greater treatment heterogeneity than expected under independence of potential outcomes and $\pi_2 < 0.22$ represents lower treatment heterogeneity than expected under independence. The confidence interval for (0, 0.46) overlaps values both below and above the value expected under independence.

6 Discussion

In this paper we presented bounds for a causal treatment effect heterogeneity parameter ($\pi_2 =$ the probably of a treatment failure and a control success). We also provided estimators for these bounds and derived formulae for the variances of the estimators and confidence intervals for each bound. Our inference is limited to bounds since π_2 is not identifiable and thus a valid point estimator is not available under the present model. Nevertheless, depending on their tightness, such bounds can provide information on plausible values of π_2. Such as assessment of treatment heterogeneity can provide insights as to underlying modifying variables and the potential for more optimal treatment strategies tailored to individual characteristics. A more complete elucidation of the practical use of information about treatment heterogeneity awaits future research.

We showed, using data from a renal clinical trial (MDRD), that the use of blocks can tighten bounds on π_2 relative to the simple bounds obtained without the use of blocks. However, the blocked data bounds for π_2 from the MDRD study data were still rather wide, leaving us with a wide range of possible levels of treatment heterogeneity. It may be that higher within-block correlations than that obtained in the MDRD data would be needed to get much narrower bounds. We found (not shown above) that varying the number of blocks based on the present covariates did not substantially affect bound width. Further, scientific conclusions should also take into account variation in the bound estimates. We thus provided a confidence interval for π_2 using the confidence intervals for the lower and upper bounds. This incorporation of variation, of course, further widens bounds and increases the challenge of obtaining practically useful information about treatment heterogeneity.

Previous bounds were available only for matched pairs data. The present method generalizes the application to general size blocks. However, the estimation of the variances of bound estimators (Section 4) does require a minimum number of subjects per block – namely, we require at least four subjects on each treatment for each block. The applicability of our method is further broadened by the fact that blocks may be constructed post-randomization.

Research by Mascha (2005) has examined the effects of key conditions (in particular, the marginal success probabilities, the within-block correlations, and the number of blocks) on the widths of the bounds for π_2. He has also explored (in work yet to be published) possible improvements on the bounds and confidence interval estimates, and has further showed that π_2 (alternatively, the correlation between the two potential outcomes) can be estimated under certain model assumptions. We are currently pursuing methods that will allow estimation of π_2 under more minimal model assumptions.

Finally, we have emphasized – as have others – the importance of gaining insights into treatment effect heterogeneity and ultimately using such information to improve treatment strategies. In light of this, it is important to study what present statistical tools have to offer as well as pursue improved techniques that will enhance our ability to reveal treatment effect heterogeneity.

Appendix

Derivation of (5)

For any two subjects, u_1 and u_2, in the same block, let $m_{jk} = P(u_1 \in \Pi_j \mid u_2 \in \Pi_k)$. Note that, assuming exchangeability of subjects within a block, we have $\pi_k m_{jk} = \pi_j m_{kj}$. Also let (j, k) denote the event that $u_1 \in \Pi_j$ and $u_2 \in \Pi_k$. We have,

$$\begin{aligned} g_2 &= P(Y_T(u_1) = 0, Y_C(u_2) = 1 \mid u_1, u_2 \text{ in same block}) \\ &= P((1,2),(2,2),(1,4),(2,4) \mid u_1, u_2 \text{ in same block}) \\ &= \pi_2 m_{12} + \pi_2 m_{22} + \pi_1 m_{41} + \pi_2(1 - m_{12} - m_{22} - m_{32}) \\ &= \pi_2 + \pi_1 m_{14} - \pi_2 m_{32}\,. \end{aligned}$$

Thus,

$$\begin{aligned} \pi_2 &= g_2 + \pi_2 m_{32} - \pi_1 m_{41} \\ &\le g_2 + \pi_2 m_{32} + \pi_1 m_{41} \\ &\le \begin{cases} g_2 + \pi_2 m_{32} + \pi_1 m_{41} + \pi_2 m_{42} + \pi_1 m_{31} \\ g_2 + \pi_2 m_{32} + \pi_1 m_{41} + \pi_3 m_{43} + \pi_1 m_{21} \end{cases}. \end{aligned}$$

or

$$\pi_2 \le g_2 + \min\,(p_T - h_T, p_C - h_C) \equiv U_B\,.$$

A similar derivation provides $L_B \equiv g_2 - \min\,(p_T - h_T, p_C - h_C)$.

Derivation of $V(\hat{h}_T)$

A sketch of this derivation is provided as an example of the approach used throughout for parameters involved in $V(\hat{U}_B)$ and $V(\hat{L}_B)$. The numerator of $\hat{h}_T$ (as given in Section 3.2) may be written as 1/2 times:

$$\sum_{j=1}^{m} \sum_{i_2 > i_1}^{n_{Tj}} \sum_{i_1=1}^{n_{Tj}-1} y_{Ti_1j} y_{Ti_2j}$$

where $y_{hij} = 1$ if subject i in block j observed on treatment h has a success; $y_{hij} = 0$ if this subject has a failure. The variance of $\hat{h}_T$ is obtained using the basic probability theory identity for the variance of a sum, and noting that the denominator of $\hat{h}_T$ is considered as fixed. For a given j, we have $V(y_{Ti_1j} y_{Ti_2j}) = h_T(1 - h_T)$ for each of the $\binom{n_{Tj}}{2} = \frac{n_{Tj}(n_{Tj} - 1)}{2}$ product terms. Also, for given j, we obtain the covariance for products that do not share an index: $C(y_{Ti_1j} y_{Ti_2j}, y_{Ti_3j} y_{Ti_4j}) = h_{T2} - h_T^2$ and the covariance for products where an index is shared: $C(y_{Ti_1j} y_{Ti_2j}, y_{Ti_3j} y_{Ti_3j}) = h_{TT} - h_T^2$. There are $\frac{1}{2}\binom{n_{Tj}}{2}\binom{n_{Tj} - 2}{2} = n_{T22j}$ of the former terms and $\left(\begin{array}{c}\binom{n_{Tj}}{2} \\ 2\end{array}\right) - n_{T22j} = n_{T4j} - n_{T22j}$ of the latter terms. The variance of the sum of products within block j is thus, after some manipulation, $h_T(1 - h_T)\, n_{T2j} + 2[(h_{T2} - h_{TT})\, n_{T22j} + (h_{TT} - h_T^2)\, n_{T4j}]$. Summing over j and multiplying by the constant term $\left(\left(\frac{1}{n_{T2\bullet}}\right)^2\right)$ gives $V(\hat{h}_T)$ as provided in Section 4.

Acknowledgements The authors thank Hari K. Iyer for early discussions that helped lead to the development of this work. The authors also thank Gerald Beck, Ph.D. and Tom Greene, Ph.D. at The Cleveland Clinic Foundation for use of the MDRD data, and the associate editor and referee for helpful comments and suggestions.

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association* **91**, 444–455.

Berger, V. W., Rezvani, A., and Makarewicz, V. A. (2003). Direct effect on validity of response run-in selection in clinical trials. *Controlled Clinical Trials* **24**, 156–166.

CAST Investigators (1989). Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infrction. *New England Journal of Medicine* **321**, 406–412.

Gadbury, G. L. and Iyer, H. K. (2000). Unit-treatment interaction and its practical consequences. *Biometrics* **56**, 882–885.

Gadbury, G. L., Iyer, H. K., and Allison, D. (2001). Evaluating subject-treatment interaction when comparing two treatments. *Journal of Biopharmaceutical Statistics* **11**(4), 313–333.

Gadbury, G. L., Iyer, H. K., and Albert, J. M. (2004). Individual treatment effects in randomized trials with binary outcomes. *Journal of Statistical Planning and Inference* **121**, 163–174.

Goetghebeur, E. and Molenberghs, G. (1996). Causal inference in a placebo-controlled clinical trial with binary outcome and ordered compliance. *Journal of the American Statistical Association* **91**, 928–934.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association* **81**, 945–960.

Longford, N. J. (1999). Selection bias and treatment heterogeneity in clinical trials. *Statistics in Medicine* **18**, 1467–1474.

Mascha, E. J. (2005). Assessing treatment effect heterogeneity for binary outcomes. Ph.D. Thesis, Case Western Reserve University.

Modification of Diet in Renal Disease (MDRD) Study Group. (Prepared by Levey, A. S., Beck, G. J., Cagguila, A. W., Greene, T., Kusek. J. W., Striker, G. E., Klahr, S.). Trends toward a beneficial effect of a low protein diet during additional follow-up in the Modification of Diet in Renal Disease Study. XXXIInd Congress of the European Renal Association, European Dialysis and Transplant Association, Athens, Greece, June 11–14, 1995.

Neter, J., Wasserman, W., and Kutner, M. H. (1985). *Applied Linear Statistical Models*. Irwin, Homewood, Illinois.

Prystowsky E. N., Katz, A., and Knilans, T. K. (1990). Ventricular arrhythmias: risk stratification and approach to therapy after the Cardiac Arrhythmia Suppression Trial (CAST). *Pacing Clinical Electrophysiology* **13**, 1480–1487.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)* **45**, 212–218.

Rosenbaum, P. R. and Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70**, 41–55.

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

Rubin, D. B. (1980). Comment on Basu's randomization analysis of experimental data. *Journal of the American Statistical Association* **75**, 591–593.

Schall, R. (1995). Assessment of individual and population bioequivalence using the probability that bioavailabilities are similar. *Biometrics* **51**, 615–626.

Sheiner, L. B. (1992). Bioequivalence revisited. *Statistics in Medicine* **11**, 1777–1788.