Estimating the Magnitude of Interaction

by

Dennis D. Boos, Cavell Brownie, and Jie Zhang
Department of Statistics, North Carolina State University
Raleigh, NC 27695-8203

Institute of Statistics Mimeo Series No. 2285

August 8, 1996

Estimating the Magnitude of Interaction

Summary

Qualitative interaction (Peto, 1982) is a useful concept, but its practical value is limited because its presence is hard to detect. This difficulty arises from trying to detect a treatment effect of opposite direction to the main effect in a small number of subgroups. Simulation results are presented which reinforce this point. A new measure of interaction magnitude, IM, is introduced and proposed as an alternative way to think about important interaction between treatments and patient subgroups. Simulation results show that comparing a lower bound for IM to a specific cutoff provides greater power to detect important interaction than do the tests for qualitative interaction.

Key words: interaction magnitude, qualitative interaction, clinical trials

1 Introduction

The primary goal of most clinical trials is to develop a reliable conclusion concerning the relative efficacy of two treatments, often a drug and a placebo. In order for the conclusion to be broad based, and to achieve reasonable sample sizes, the study design often includes a number of patient subgroups, corresponding to different clinical centers or to patient subsets defined by prognostic factors. As a result, some variation in the estimates of treatment effect among subgroups is expected. (The term "treatment effect" denotes a measure of the difference in efficacy of the two treatments such as the difference in mean responses or the ratio of success rates.) Assessment of the variation in treatment effect among subgroups is important for several reasons. For example, the pattern of variation may lead to specific hypotheses concerning the relationship between efficacy and certain patient charateristics. More often, especially in Phase III multicenter trials, the task is to decide whether the variation in effect is inconsequential, or whether it is important enough to prevent making conclusions about average treatment benefit.

The conventional approach to assessing variation of treatment effect among subgroups is to test the hypothesis of no interaction between the treatment and subgroup effects. The procedures used to test for treatment by subgroup interaction are reviewed in Simon (1982) for several different measures of treatment effect, including the difference in mean response, the log odds ratio and the log relative hazard. If the test for interaction fails to be significant, then conclusions concerning average treatment effect can be made, but if the hypothesis of no interaction is rejected, convention precludes making statements about average treatment benefit.

An obvious problem with reporting conclusions based on the outcome of conventional tests for interaction is that a statistically significant interaction does not necessarily correspond to a clinically important interaction. In large trials power to detect interaction will be high and minor variation in treatment effect, though not of medical importance, will be declared statistically significant. To focus attention on interaction that is of clinical importance, Peto (1982) used the term qualitative interaction to refer to the situation where one treatment is superior for some subgroups and the other treatment is better for the remaining groups. In contrast, quantitative interaction was

used by Peto to represent variation in magnitude, but not in direction, of treatment effect. Clearly, a statement about average treatment benefit cannot be made if substantial qualitative interaction is present.

The statistical problem of determining whether observed variability in treatment effect represents qualitative interaction was addressed by Gail and Simon (1985). These authors, who use the more visual terms "crossover and noncrossover" interaction for Peto's "qualitative and quantitative" interaction; developed a normal theory likelihood ratio test to detect the presence of qualitative interaction. Other tests followed including those of Berger (1989, Section 6), Zelterman (1990), the pushback tests of Ciminera *et al.* (1993), and the range test of Piantadosi and Gail (1993). Unfortunately, however, these tests perform poorly; the pushback tests can be liberal, and results in Piantadosi and Gail (1993) indicate that unreasonably large sample sizes may be needed to obtain adequate power with both the range and Gail and Simon tests.

We have derived an upper bound for the power of the Gail and Simon test which helps to illustrate why the power will be low in a typical clinical trial. Consider a multicenter trial with ten centers and equal sample sizes in each center chosen so that the analysis of variance test for H_0 : "No treatment main effect" has power 80% to detect an average treatment effect of δ. If eight of the centers have treatment effect 1.5δ and two of the centers have treatment effect $-\delta$, then we can show that power of the Gail and Simon test for qualitative interaction must be less than the power of the best test of H_0 based on only the two "negative direction" centers. (This argument is made formal in the Appendix using the Neyman-Pearson Lemma.) Since sample sizes were chosen so that power is 80% for detecting an average treatment effect δ based on 10 centers, it follows that a test based on only two centers will have power which is considerably less than 80%.

A practical consequence of the low power of the test for qualitative interaction is that treatment by subgroup interaction, though present, may no longer appear to be a "problem" in clinical trials. To explain further, suppose that for a particular drug there is substantial variation in treatment benefit across subgroups, and data from a trial involving this drug and a placebo are analysed using standard tests for treatment main effect and interaction. A likely outcome is that the standard test for interaction is

significant, and the researchers proceed to determine whether interaction is "important" using a test for qualitative interaction. As a result of low power, this test fails to detect qualitative interaction, leading the researchers to conclude erroneously that variation in treatment effect is unimportant. In other words, the pendulum has swung back from a tendency to overreact to low p-values for the conventional test for interaction to a situation where medically important interaction may be overlooked.

Our goal in this article is not to criticize the concept of qualitative interaction (which we find appealing), nor the tests designed to detect its presence. Rather we want to emphasize how hard it is to detect qualitative interaction, and also to suggest that there are situations where there is no crossover of treatment effects but where interaction is nevertheless of clinical importance.

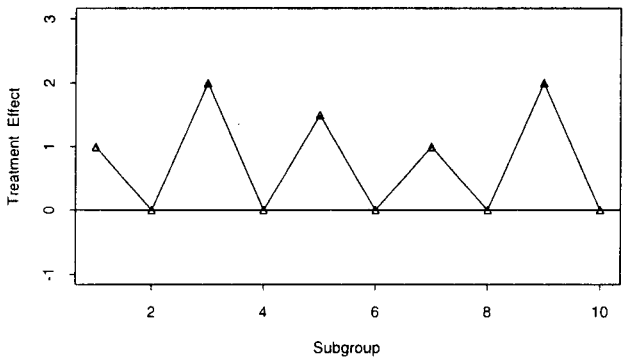

Figure 1: An illustration of quantitative but not qualitative interaction.

Figure 1 displays the treatment effects δ_i for a hypothetical situation where quantitative, but not qualitative, interaction is present. Note that for five of the ten subgroups the treatment has no benefit ($\delta_i = 0$), while in the other five subgroups the treatment has a positive effect ($\delta_i > 0$). The average treatment effect is positive and there is no crossover of effect, but the degree of interaction is nonnegigible and suggests further investigation to discover why there is no benefit, on average, for a substantial part of the patient population. This example motivates our main objective, which is to propose a new measure of the importance of treatment by subgroup interaction. This measure, which we call the Interaction Magnitude (IM), reflects the variation in treatment effect

relative to the average effect.

In Section 2 we review tests for qualitative interaction and present an upper bound for power of the Gail and Simon test. Also in Section 2, we provide simulation results for the performance of several tests for qualititative interaction. In Section 3 we introduce the IM measure and and an associated lower confidence bound, and simulation results comparing IM and the tests for qualitative interaction are reported in Section 4. Sections 5 and 6 contain an example and conclusions, respectively.

2 Simulation Results for Tests for Qualitative Interaction

Let $\{D_i\}$ denote the estimate of treatment effect in the ith subgroup, $i = 1, 2, \ldots, q$. The tests for qualitative interaction assume that the $\{D_i\}$ are independent and normally distributed with means $\{\delta_i\}$ and known variances $\{\sigma_i^2\}$. Typically, the D_i are obtained in a clinical trial with two treatments in q subgroups from data Y_{ijk} described by the linear model

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, q, \quad j = 1, 2, \quad \text{and} \quad k = 1, \ldots, n_{ij},$$

where μ_{ij} is the mean response from the ith subgroup and the jth treatment level, and the errors ϵ_{ijk} are iid $N(0, \sigma^2)$. The treatment effect in each subgroup is $\delta_i = \mu_{i1} - \mu_{i2}$, which is estimated by $D_i = \overline{Y}_{i1.} - \overline{Y}_{i2.}$, where the variance of D_i is $\sigma_i^2 = \sigma^2(1/n_{i1} + 1/n_{i2})$. As in Gail and Simon (1985), we assume here that σ^2 is known in order to simplify exposition. In order to carry out the test, σ^2 is replaced by a consistent estimate. Also, D_i can be any measure of treatment effect which is approximately normally distributed such as the log odds ratio or log relative hazard.

The null hypothesis of no qualitative interaction is

$$H_0: \quad \Delta \in O^+ \cup O^-,$$

where $\Delta^T = (\delta_1, \ldots, \delta_q)$, $O^+ = \{\Delta: \ \delta_i \geq 0 \text{ for all } i\}$ and $O^- = \{\Delta: \ \delta_i \leq 0 \text{ for all } i\}$.

The likelihood ratio test of Gail and Simon (1985) rejects H_0 if both

$$Q^- \equiv \sum (D_i^2/\sigma_i^2) I(D_i > 0) > C_\alpha \quad \text{and} \quad Q^+ \equiv \sum (D_i^2/\sigma_i^2) I(D_i < 0) > C_\alpha,$$

where $I(D_i > 0) = 1$ if $D_i > 0$ and 0 otherwise, and $I(D_i < 0) = 1$ if $D_i < 0$ and 0 otherwise. The critical value C_α, which may be obtained from Table 1 of Gail and Simon (1985), is chosen so that the test has level at most α under any Δ for which there is no crossover.

In the Appendix, assuming equal sample sizes for each of q subgroups, we obtain an upper bound for the power of the Gail and Simon test under alternatives where the average treatment effect is $\bar{\delta}$, q_1 subgroups have the same negative effect $-c_1\bar{\delta}$, and $q - q_1$ subgroups have the same positive effect $c_2\bar{\delta}$ with $c_2 = (1 + c_1 q_1/q)/(1 - q_1/q)$. If sample sizes are chosen to give power of .80 for a test at level .05 to detect the treatment main effect, this upper bound is

$$P\left(Z < -\Phi^{-1}(.95) + \frac{\sqrt{q_1}c_1}{\sqrt{q}}\left[\Phi^{-1}(.975) - \Phi^{-1}(1 - .80)\right]\right). \tag{1}$$

For $q = 10$, $q_1 = 2$, and $c_1 = 1$, this bound is .35, and changing q_1 to 1 lowers the bound to .22. With $q = 10$, $q_1 = 4$, and $c_1 = .5$, the bound is also .22. Each of these cases illustrates a situation where substantial crossover is likely to go undetected by the Gail and Simon test given sample sizes typical of most clinical trials.

The range test of Piantadosi and Gail (1993) rejects H_0 at level α if both

$$max\{D_i/\sigma_i\} > C'_\alpha \quad \text{and} \quad min\{D_i/\sigma_i\} < -C'_\alpha.$$

where C'_α may be obtained from Table 1 of Piantadosi and Gail (1993). As in Gail and Simon (1985) the critical values C'_α were obtained assuming normality of the D_i and σ^2 known.

Ciminera *et al.* (1993) proposed several procedures to test H_0, but as these "push-back" procedures are more complicated to describe, we refer readers to that source.

The likelihood ratio test, the range test, and the pushback procedure in Section 2.2 of Ciminera *et al.* were compared by simulation for 12 situations corresponding to different patterns of interaction. For comparison with the tests for qualitative interaction, the normal theory tests for treatment main effect and treatment by subgroup interaction

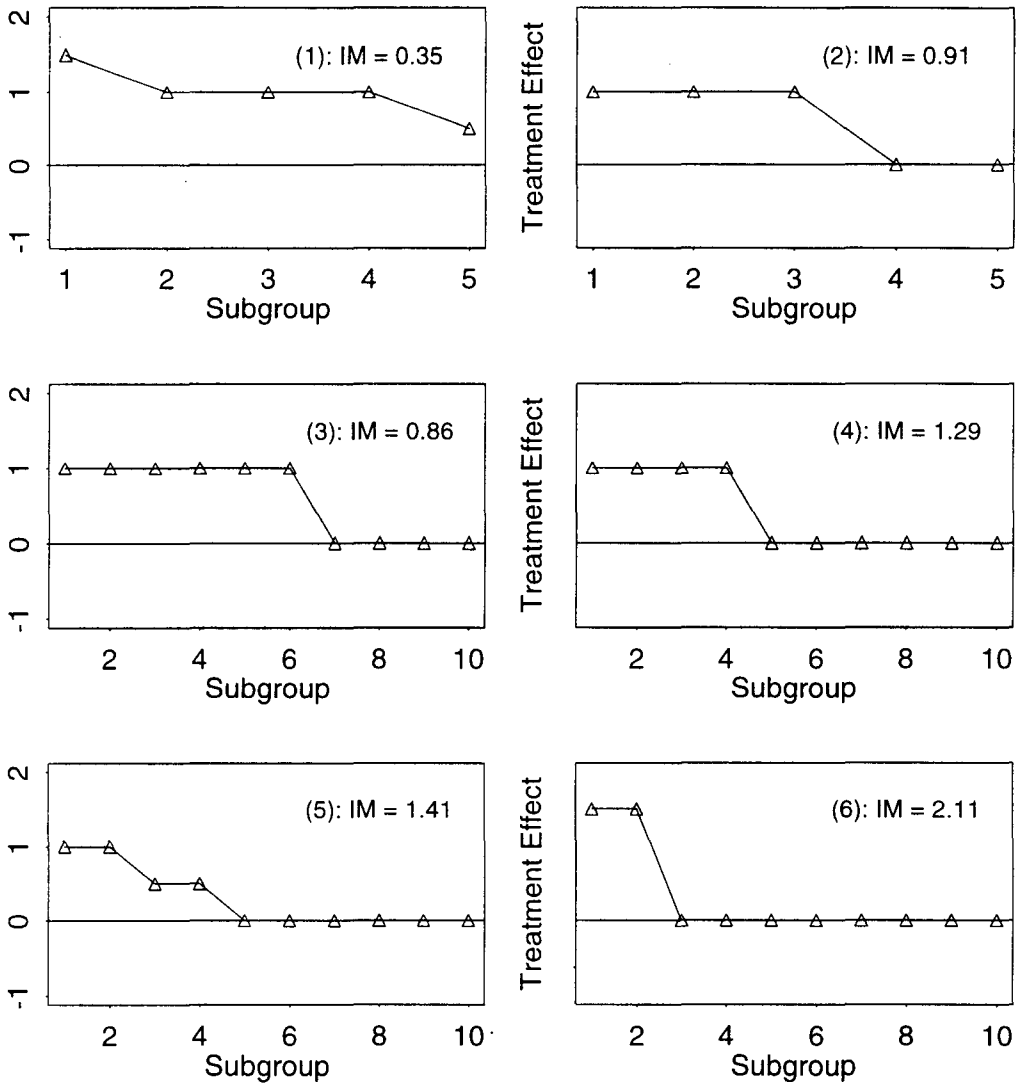

Figure 2: Treatment effect versus subgroup for situations with only quantitative interaction.

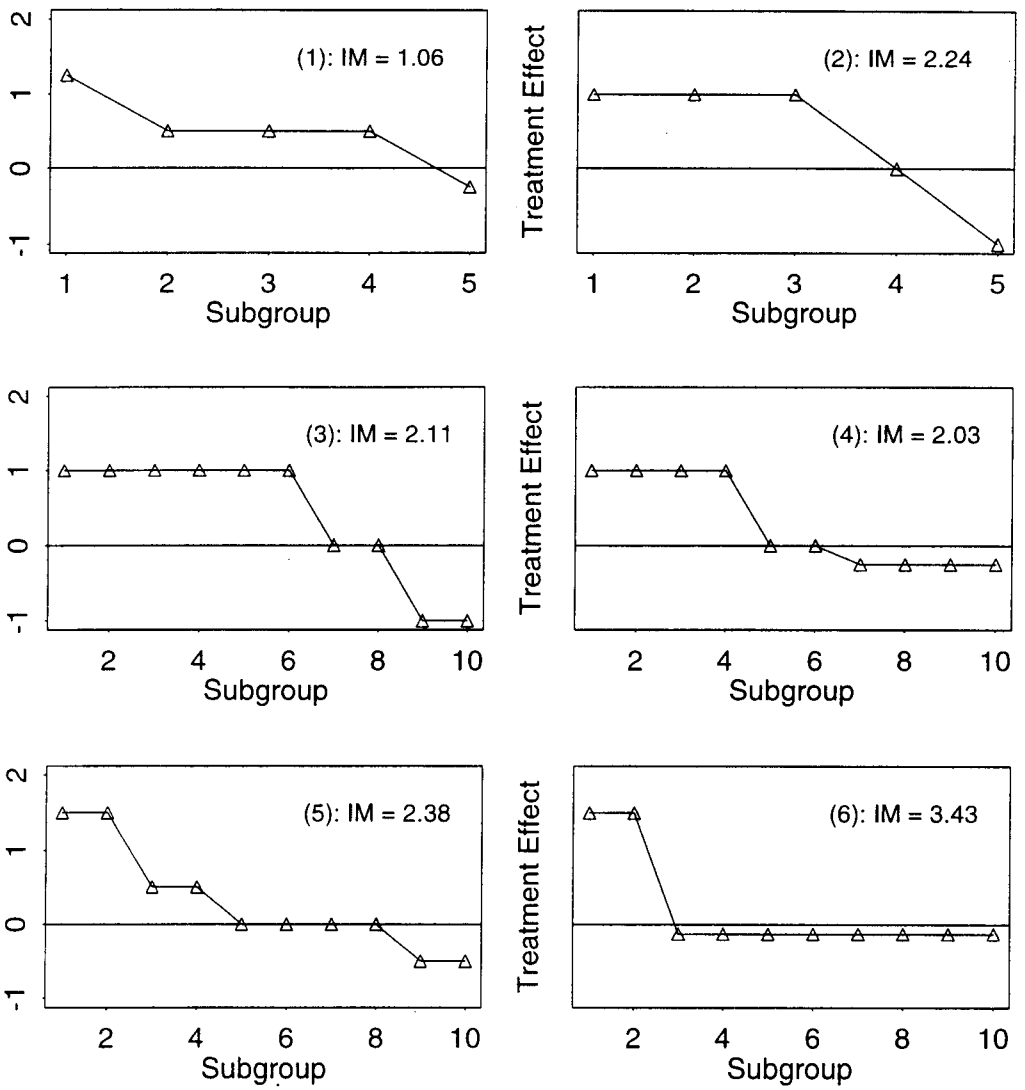

Figure 3: Treatment effect versus subgroup for situations with qualitative interaction.

were included. For simplicity, σ^2 was assumed known so that the normal theory test statistics are, respectively,

$$T_1 = \frac{(\sum_{i=1}^{q} D_i)^2}{\sum_{i=1}^{q} \sigma_i^2} \text{ and } T_2 = \sum_{i=1}^{q} \frac{(D_i - \hat{D})^2}{\sigma_i^2}$$

with $\hat{D} = [\sum_{i=1}^{q} D_i/\sigma_i^2]/[\sum_{i=1}^{q} 1/\sigma_i^2]$. Under the hypothesis of no treatment main effect, T_1 has a central chi-squared distribution with 1 degree of freedom, and under the hypothesis of no interaction, T_2 has a central chi squared distribution with $q - 1$ degrees of freedom.

The 12 cases studied are represented in Figures 2 and 3 by plotting the (ordered) values of the treatment effects δ_i. Note that in each of the 12 cases there is a positive treatment main effect ($\bar{\delta} > 0$) and also treatment by subgroup interaction. For the six cases in Figure 2 interaction is of the noncrossover type, and the hypothesis of no qualitative interaction holds, while the six cases in Figure 3 display qualitative interaction. These are of course hypothetical situations, but with the exception of Figure 2.1, they each represent a situation where an average treatment effect is of questionable value as a single summary statistic due to the presence of nonnegligible interaction.

In cases 2 and 3 of Figure 3 there are positive and negative effects of equal magnitude, similar to patterns studied in Piantadosi and Gail (1993). As qualitative interaction will not always involve effects of equal magnitude in both directions, the other four cases have smaller effects in the negative direction than in the positive direction. In both Figures 2 and 3, cases 1 to 6 were chosen to represent treatment main effects of decreasing magnitude. Also, cases with the same number in these figures display roughly similar interaction patterns (though without crossover in Figure 2).

For each of the 12 cases, 1000 Monte Carlo replications were generated assuming a balanced study design with $n = 100/q$ subjects per treatment per subgroup. In each replicate, treatment estimates $D_i, i = 1, \ldots, q$, were generated as independent normal variates with means δ_i and variances $\sigma_i^2 = 2/n$ using the S-Plus function "rnorm." Results are reported in Table 1 as empirical rejection rates for a nominal significance level of .05 for the tests for main effect and interaction as well as for the Gail and Simon likelihood ratio test (LR), the range test and the pushback test (PB).

Results for T_1 and T_2 in Table 1 serve mainly to calibrate the magnitude of effects. Thus for $\bar{\delta} \geq .4$, with 200 subjects T_1 has good power to detect a treatment main effect. Also, except for the noncrossover cases 1 and 5, T_2 has power of at least .6. Of greater interest are the LR, range and PB tests for qualitative interaction. Results for the first six cases represent performance of these tests in the null situation. Note that cases 4 to 6 show that the PB procedure can be decidedly liberal, e.g., in situations where some subgroups have large positive effects and there are many other subgroups with effects close to zero. In contrast, the LR and range tests are conservative.

Rejection rates for the second set of six cases in Table 1 represent power to detect qualitative interaction, though results for the PB test are not meaningful because of the liberal behavior of this test under the null. It is evident that the LR and range tests have low power except in cases 2 and 3. These tests are able to detect qualitative interaction when there are large effects in both directions, but have poor power when a large effect in one direction is balanced by many small effects of opposite sign. Case 1 is particularly alarming since main effect and interaction in the conventional sense will be detected but qualitative interaction will be missed. In addition, our results suggest that power and sample size calculations in Piantadosi and Gail (1993) may be optimistic because they relate to situations like cases 2 and 3 rather than to the cases where power is poor for both the LR and range tests.

3 Interaction Magnitude

It is clear from results in Section 2 that clinically important interaction may not be detected by the tests for qualitative interaction. On the other hand, the conventional tests for interaction will frequently be significant when interaction is not clinically meaningful. As an alternative, we propose a measure which attempts to quantify the seriousness of interaction. This measure is the interaction magnitude defined by

$$IM = \frac{\sqrt{\frac{1}{q-1}\sum_{i=1}^{q}(\delta_i - \bar{\delta})^2}}{|\bar{\delta}|},$$

Table 1: Power estimates for interaction described in Figures 2 and 3

Case No.	T_1	T_2	LR	Range	PB	$\bar{\delta}$	IM
	Only Quantitative Interaction (Fig. 2)						
1	1.00	0.40	0.00	0.00	0.00	1.0	0.35
2	0.99	0.80	0.01	0.02	0.03	0.6	0.91
3	0.99	0.66	0.01	0.02	0.06	0.6	0.86
4	0.79	0.64	0.02	0.02	0.13	0.4	1.29
5	0.56	0.45	0.02	0.03	0.17	0.3	1.41
6	0.58	0.86	0.03	0.04	0.31	0.3	2.11
	Qualitative Interaction (Fig. 3)						
1	0.94	0.77	0.04	0.08	0.07	0.5	1.06
2	0.81	1.00	0.76	0.82	0.81	0.4	2.24
3	0.80	0.99	0.54	0.58	0.81	0.4	2.11
4	0.56	0.83	0.10	0.09	0.40	0.3	2.03
5	0.56	0.94	0.16	0.19	0.48	0.3	2.38
6	0.30	0.92	0.13	0.09	0.58	0.2	3.43

T_1: test for no treatment effect, T_2: test for no quantitative interaction.

LR, Range, PB=Pushback: tests for no qualitative interaction.

IM: definition of IM is given in Section 3.

where $\bar{\delta} = q^{-1} \sum_{i=1}^{q} \delta_i$. Note that IM is just the coefficient of variation of the treatment effects $\{\delta_i\}$. In other words, IM reflects variation in the treatment effects among subgroups relative to the average treatment effect $\bar{\delta}$. Thus, if IM is small, the average treatment effect $\bar{\delta}$ is a good summary of overall benefit since effects vary little among the individual subgroups. On the other hand, large values of IM indicate that variation in the δ_i is considerable, or that $\bar{\delta}$ is close to zero. In either case, the treatment benefit is questionable, and reporting an average treatment effect is not appropriate.

For situations where there is a treatment main effect ($\bar{\delta}$ is not close to zero), we need guidelines concerning what values of IM represent non-negligible interaction. That is, what cut-off IM^* suggests that values exceeding IM^* indicate the presence of medically important interaction? One approach to providing such a cutoff is based on assuming

that the $\delta_1, \ldots, \delta_q$ are iid $N(\overline{\delta}, \sigma_\delta^2)$ so that IM will estimate $\sigma_\delta/\overline{\delta}$. Then the proportion of treatment effects δ_i which are less than a fraction a of the average treatment effect is

$$P(\delta_i \leq a\overline{\delta}) = P(Z \leq (a-1)/IM) = \Phi((a-1)/IM). \tag{2}$$

where Z has a standard normal distribution with distribution function Φ. Table 2 summarizes a few calculations based on (2). For example, if IM is about 1 and $\overline{\delta} > 0$, then we would expect a negative treatment effect in about 16% of the subgroups. If $IM = .5$ then we would expect a negative treatment effect in just 2% of the sugroups.

Table 2: Proportion of subgroup means $\leq a\overline{\delta}$

	Interaction Magnitude ($IM = \sigma/\overline{\delta}$)					
Fraction	0.5	1	1.5	2	2.5	∞
$a = 0$	0.02	0.16	0.25	0.31	0.34	0.5
$a = 1/3$	0.09	0.25	0.33	0.37	0.39	0.5
$a = 1/2$	0.16	0.31	0.37	0.40	0.42	0.5

The interaction patterns in Figures 2 and 3 should also help to calibrate the values of IM with respect to importance of interaction. In addition, we estimated IM (using the method in Section 3.1) for examples in Gail and Simon (1985) and in Ciminera et al. (1993). For Table 2 in Gail and Simon, treatment effect was the difference in proportion of disease-free patients after 3 years, and the four subgroups yielded estimates D_i (and standard errors) equal to .163 (.0788), −.114 (.0689), −.047 (.0614) and −.151 (.0547), with $\overline{D} = -.04$. The estimated value of IM is .41.

In Table 3 of that paper, where treatment effect is obtained as the log relative hazard, estimates D_i (and standard errors) are .531 (.208), −.266 (.182), −.030 (.206) and −.724 (.207), giving $\overline{D} = -.122$ and $\widehat{IM} = 1.93$. Ignoring variability associated with these IM estimates, using a cut-off of .5 suggests that interaction is not serious when effects are measured in terms of differences in rates, but that interaction is more important if treatment effects are expressed as log relative hazards. Similar conclusions were reached using the LR test, but the pushback test found important interaction on both scales (Ciminera et al., 1993).

A third example from Table 1 in Ciminera et al. (1993) yields values of D_i between -7.0 and 12.9, where D_i is the difference in mean response for a placebo and drug. We obtain $\overline{D} = 5.6$ and $\widehat{IM} = .81$ suggesting interaction may be of concern. Ciminera et al. (1993) report that there is "no substantial indication of qualitative interaction."

3.1 Estimation of IM

We propose the following estimator $\widehat{IM}$ for IM:

$$\widehat{IM} = \frac{\hat{\sigma}_\delta}{\left|\overline{D}\right|}$$

with $\overline{D} = q^{-1}\sum_{i=1}^{q} D_i$, and

$$\hat{\sigma}_\delta^2 = \begin{cases} \frac{1}{q-1}\sum_{i=1}^{q}(D_i - \overline{D})^2 - \frac{1}{q}\sum_{i=1}^{q}\hat{\sigma}_i^2 & \text{if } \frac{1}{q-1}\sum_{i=1}^{q}(D_i - \overline{D})^2 > \frac{1}{q}\sum_{i=1}^{q}\hat{\sigma}_i^2 \\ 0 & \text{otherwise} \end{cases}$$

If we assume that D_i is an unbiased estimator of δ_i and $\hat{\sigma}_i^2$ is an unbiased estimator of $Var(D_i)$, then it is easy to show that

$$E\left[\frac{1}{q-1}\sum_{i=1}^{q}(D_i - \overline{D})^2 - \frac{1}{q}\sum_{i=1}^{q}\hat{\sigma}_i^2\right] = \frac{1}{q-1}\sum_{i=1}^{q}(\delta_i - \overline{\delta})^2.$$

However $\hat{\sigma}_\delta^2$ is positively biased as an estimator of σ_δ^2 because of the truncation at 0.

When $\overline{\delta}$ is close to zero, estimation of IM is difficult because small values of $\left|\overline{D}\right|$ will occur resulting in frequent large estimates of IM. However, $\overline{\delta}$ is close to zero only when 1) there is no treatment effect, or 2) treatment effects have opposite directions in different subgroups, and none of the directions dominate. These correspond to situations where the treatment main effect is close to zero, so that there is no strong need to evaluate the importance of interaction (or to estimate IM).

Analytical methods for obtaining interval estimates for IM are tedious and will be sensitive to distributional assumptions for the effect estimates D_i and related variance estimates. Instead, we propose a confidence interval for IM as follows. First we use the standard jackknife to compute an estimate of the covariance matrix of $(\hat{\sigma}_\delta, \left|\overline{D}\right|)$, say,

$$\begin{pmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{pmatrix}.$$

Now let

$$\begin{aligned} a &= \overline{D}^2 - Z_{1-\alpha}^2 V_{22}, \\ b &= 2\left|\overline{D}\right| \hat{\sigma}_\delta - 2Z_{1-\alpha}^2 V_{12}, \\ c &= \hat{\sigma}_\delta^2 - Z_{1-\alpha}^2 V_{11}, \end{aligned}$$

where $Z_{1-\alpha}$ is the $1-\alpha$ quantile of a standard normal distribution.

Then by Fieller's Theorem (Casella and Berger, 1990, p. 459), an approximate $100(1-2\alpha)\%$ confidence interval for IM is

$$\left(\widehat{IM}_{L,1-\alpha}, \widehat{IM}_{U,1-\alpha}\right) = \left(\frac{b - \sqrt{b^2 - 4ac}}{2a}, \frac{b + \sqrt{b^2 - 4ac}}{2a}\right),$$

provided $b^2 - 4ac > 0$, $a > 0$, and $c > 0$. Several special cases where these inequalities do not all hold need to be considered separately. If $b^2 - 4ac > 0$, $a > 0$ but $c < 0$ then $\widehat{IM}_{L,1-\alpha}$, the lower endpoint of the interval, is defined to be 0. If $b^2 - 4ac > 0$, $c > 0$, but $a < 0$, then $\widehat{IM}_{U,1-\alpha} = \infty$. If $b^2 - 4ac < 0$, or if both $a < 0$ and $c < 0$, then the interval is $[0, \infty)$.

To demonstrate the presence of important interaction, we use the left endpoint of the interval as a $1-\alpha$ lower bound for IM. This leads to the following level α test to detect interaction:

1. Use (1) and/or scientific understanding to specify a cutoff value IM^*. We suggest .5 as a possible value for IM^*.

2. Compute $\widehat{IM}$ and the $(1-\alpha)\%$ lower bound $\widehat{IM}_{L,1-\alpha}$ (keeping in mind the special cases above). If $\widehat{IM}_{L,1-\alpha} > IM^*$, non-negligible interaction has been found.

Of course, if a non-significant result is found for the treatment main effect, there will be no direct practical interest in computing $\widehat{IM}$ or $\widehat{IM}_{L,1-\alpha}$.

4 Simulation Results for $\widehat{IM}$

We have suggested that the tests for qualitative interaction generally have low power. Here we examine the performance of the estimator $\widehat{IM}$ and of a test for interaction which is based on comparing the lower bound $\widehat{IM}_{L,.95}$ to the cutoff value $IM^* = .5$. Performance of this test is compared with that for the LR and range tests for qualitative interaction, but the PB test is omitted because of its unsatisfactory behavior in certain null situations.

Cases 1-6 of Figure 3 were chosen for further simulation study. In order to carry out the jacknife procedure values for individual subjects are needed, and so for each Monte Carlo sample, data were generated for the 200 subjects according to the linear model given in Section 2,

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk},$$

where the errors ϵ_{ijk} are iid $N(0,1)$, $\mu_{i1} = \delta_i$ and $\mu_{i2} = 0$, with the δ_i as in Figure 3. One thousand Monte Carlo samples were generated for each case. For comparison with the IM procedure, the LR and range tests for qualitative interaction were carried out with σ^2 assumed unknown. The conventional analysis of variance F tests for treatment main effect (T_1) and quantitative interaction (T_2) were also included for comparison. The values of D_i and $\hat{\sigma}_i^2$, which are needed for computing the LR test, the range test, and $\widehat{IM}$, are calculated as follows:

$$D_i = \overline{Y}_{i1.} - \overline{Y}_{i2.} \text{ and } \hat{\sigma}_i^2 = \hat{\sigma}^2(\frac{1}{n_{i1}} + \frac{1}{n_{i2}}),$$

where

$$\hat{\sigma}^2 = \frac{\sum_i \sum_j \sum_k (Y_{ijk} - \overline{Y}_{ij.})^2}{N - 2q}.$$

Results for the estimator $\widehat{IM}$ and empirical rejection rates for each test are presented in Table 3. Because large $\widehat{IM}$ values occur when $\overline{D}$ is close to 0, we report the median and standardized interquartile range = IQR = (sample 75th percentile − 25th percentile)/1.349 in place of the mean and standard deviation of the $\widehat{IM}$ values. The

Table 3: Power estimates and properties of $\widehat{IM}$ for cases 1-6 of Figure 3

Case					$\widehat{IM}_{L..95}$	$\widehat{IM}_{L..95}$	$\widehat{IM}$		True
No.	T_1	T_2	LR	Range	Coverage	> 0.5	Median	IQR	IM
1	.93	.74	.03	.06	.97	.37	1.02	0.49	1.06
2	.79	1.00	.75	.82	.95	.98	2.20	0.96	2.24
3	.79	.99	.56	.57	.95	.94	2.09	0.94	2.11
4	.56	.81	.08	.09	.97	.65	1.98	1.15	2.03
5	.56	.93	.15	.15	.97	.83	2.28	1.34	2.38
6	.30	.92	.10	.09	.99	.87	3.21	2.69	3.43

T_1: test for no treatment main effect, T_2: test for no quantitative interaction.

LR, Range: tests for no qualitative interaction.

$\widehat{IM}_{L..95}$ > 0.5: test based on 95% lower confidence bound.

IQR: (Interquartile Range)/1.349.

median value of $\widehat{IM}$ slightly underestimates the true value because the sampling distribution is right skewed. As expected, the variability in the estimates $\widehat{IM}$ increases as the value of $\bar{\delta}$ (see Table 1) decreases. Even though the estimator $\widehat{IM}$ can produce extreme values with nonnegligible probability, coverage of the one sided interval for IM is satisfactory (or slightly conservative) in all cases.

Comparison of estimated power in Table 3 with analogous results in Table 1 indicates that estimating σ^2 has little effect on the power of T_1, T_2, the LR and range tests, at least for the sample sizes studied. Thus we see again that both the LR and the range tests for qualitative interaction have low power except for cases 2 and 3 where there are large treatment effects in both directions. In every case the test based on comparing the lower bound $\widehat{IM}_{L..95}$ to 0.5 has substantially greater power than the LR and range tests for qualitative interaction. Only in case 1, where IM = 1.06, does the the test based on $\widehat{IM}_{L..95}$ have low power (.37). Recall that this is the case where the treatment main effect is strong and the LR and range tests have almost no chance of detecting crossover of treatment efects.

5 Example

The data reported in Table 4 are from a multicenter clinical trial designed to compare a new drug for topical treatment of psoriasis with a placebo (see Boos and Brownie, 1992). The measure of treatment effect reported in Table 4 is the difference in means (drug – placebo) for a score representing improvement. Standard errors, $\hat{\sigma}_i$, are obtained as in Section 4. The analysis of variance tests for treatment main effect and quantitative interaction are obtained using the SAS GLM procedure (Type III sums of squares) in a fixed effects model. Resulting p-values for treatment main effect and interaction are 0.0005 and 0.033, respectively. The p-values of the tests for qualitative interaction are all > 0.2. $\widehat{IM}$ is 0.84, with lower confidence bounds 0.11 (95%), 0.25 (90%) and 0.43 (80%). Therefore neither the tests for qualitative interaction nor $\widehat{IM}$ provide evidence of non-negligible interaction.

For illustration, we show the results of these tests after changing the data for Center 2 and Center 4 from $(D_i, \hat{\sigma}_i^2)$ = (0.20, 0.26) and (−0.10, 0.14) to $(D_i, \hat{\sigma}_i^2)$ = (−0.40, 0.25) and (−0.43. 0.13). The p-values for F tests for treatment main effect and interaction become 0.0038 and 0.0030, respectively. The LR and range tests for qualitative interaction are still not significant at the α = 0.2 level. In contrast, $\widehat{IM}$ is 1.41, with lower confidence bounds of 0.52 (95%), 0.68 (90%), and 0.89 (80%), indicating presence of a non-negligible interaction.

6 Conclusions

Evaluation of the interaction between the treatment effect and subgroups is a difficult problem in clinical trials. The primary objective is to decide whether or not the variation should affect the overall conclusion concerning "average" treatment benefit, but conventional analysis of variance tests do not distinguish between clinically important and unimportant interaction. The alternative approach of testing for qualitative interaction also has disadvantages. First, there are situations (Figure 2) where there is no crossover of treatment effect but where quantitative interaction is too strong to be ignored. Second, it is evident from our simulations that the power of procedures that test for qualitative

Table 4: Summary of improvement scores after treatment with placebo (Y_1) or drug (Y_2)

	No. of Patients					
Center	n_{i1}	n_{i2}	$\overline{Y}_{i1.}$	$\overline{Y}_{i2.}$	$D_i = \overline{Y}_{i2.} - \overline{Y}_{i1.}$	$\hat{\sigma}_i^2$
1	10	10	2.70	3.30	0.60	0.12
2	5	4	2.80	3.00	0.20	0.26
3	8	8	3.00	4.37	1.38	0.15
4	9	8	3.22	3.13	−0.10	0.14
5	11	11	3.18	3.18	0.00	0.11
6	7	8	2.86	3.50	0.64	0.16
7	3	3	2.67	2.33	−0.33	0.39
8	8	8	2.50	3.75	1.25	0.15
9	6	4	2.67	3.75	1.08	0.24
					$\overline{D}$ = 0.52	

interaction is often too low to detect realistic levels of qualitative interaction.

Instead of developing more powerful tests to detect qualitative interaction, we have introduced a new quantity, the interaction magnitude, or IM, which reflects the variation of treatment effect among subgroups relative to the average treatment effect. Comparing a lower $(1 - \alpha)$ bound for IM to a cutoff IM^* (e.g., $IM^* = .5$) provides a test to detect clinically important interaction. Simulation results suggest this procedure is a more reliable indicator of important interaction than existing tests.

Acknowledgement

We would like to thank Roger Berger for careful reading of the manuscript and for many helpful comments.

Appendix

Derivation of the bound (1) to the power function of the Gail and Simon (1985) test.

We assume that there are q centers with estimated treatment effects $D_1, \ldots, D_q$ which are independent and normally distributed with $Var(D_i) = \sigma_i^2$. For simplicity we further assume that all centers have the same sample sizes so that $\sigma_i^2 = \sigma_0^2$. If the sample sizes have been chosen so that the test based on T_1 of Section 2 has 80% power at $\alpha = .05$ to detect a true treatment effect $\overline{\delta} = \delta$, then

$$\begin{aligned} .80 = P(T_1 > 3.84) &= P\left(\frac{\sqrt{q}\overline{D}}{\sigma_0} < -1.96\right) + P\left(\frac{\sqrt{q}\overline{D}}{\sigma_0} > 1.96\right) \\ &= P\left(\frac{\sqrt{q}(\overline{D} - \delta + \delta)}{\sigma_0} < -1.96\right) + P\left(\frac{\sqrt{q}(\overline{D} - \delta + \delta)}{\sigma_0} > 1.96\right) \\ &= P\left(Z < -1.96 - \frac{\sqrt{q}\delta}{\sigma_0}\right) + P\left(Z > 1.96 - \frac{\sqrt{q}\delta}{\sigma_0}\right), \end{aligned}$$

where Z is a standard normal random variable. Because the first term will be very close to zero, we ignore it and solve to get

$$\frac{\delta}{\sigma_0} = \frac{1.96 - \Phi^{-1}(1 - .80)}{\sqrt{q}}.$$

The above calculation has fixed the value of δ/σ_0. Now (for simplicity) consider alternatives where $q_1 = 2$ of the centers have negative effect $-\delta$ and the other centers have positive effects. A related testing situation is

$$H_0 : \delta_1 = 0, \delta_2 = 0, \delta_i > 0, i = 3, \ldots, q;$$

versus

$$H_a : \delta_1 = \delta_2 = -\delta, \delta_i > 0, i = 3, \ldots, q;$$

By the Neyman-Pearson lemma (Casella and Berger, 1990, p. 366), the most powerful $\alpha = .05$ test in this situation rejects if

$$\sqrt{2}\frac{(D_1 + D_2)}{2\sigma_0} < -1.645.$$

Adding and subtracting $-\delta$ and substituting for δ/σ_0, leads to the power for this test:

$$P\left(Z < -1.645 + \frac{\sqrt{2}\delta}{\sigma_0}\right) = P\left(Z < -1.645 + \frac{\sqrt{2}}{\sqrt{q}}\left[1.96 - \Phi^{-1}(1 - .80)\right]\right). \quad (3)$$

Since H_0 above belongs to the null hypothesis of the Gail and Simon LR test, then the most powerful test above must, by definition, have greater power than a size $\alpha = .05$ LR test. Thus (3) is an upper bound for power of the LR test for alternatives where $\bar{\delta} > 0$ and there are 2 negative direction centers with $\delta_i = -\bar{\delta}$.

More generally, let q_1 centers have effect $-c_1\bar{\delta}$ and $q - q_1$ centers have effect $c_2\bar{\delta}$, where $c_2 = (1 + c_1 q_1/q)/(1 - q_1/q)$, so that the average treatment effect over all q centers is $\bar{\delta}$. For sample sizes chosen so that an α level test based on T_1 has power P for a constant alternative $\bar{\delta}$, the Gail and Simon (1985) LR test at level α has power bounded by

$$P\left(Z < -\Phi^{-1}(1 - \alpha) + \frac{\sqrt{q_1}c_1}{\sqrt{q}}\left[\Phi^{-1}(1 - \alpha/2) - \Phi^{-1}(1 - P)\right]\right).$$

References

Berger, R. (1989). Uniformly more powerful tests for hypotheses concerning linear inequalities and normal mean. *Journal of the American Statistical Association* 84, 192-199.

Boos, D.D. and Brownie C. (1992). A rank-based mixed model approach to multisite clinical trials. *Biometrics* 48, 61-72.

Casella, G., and Berger, R. L. (1990). *Statistical Inference*, Wadsworth & Brooks/Cole, Pacific Grove, California.

Ciminera, J., Heyse, J., Nguyen, H. and Tukey, J. (1993). Tests for qualitative treatment-by-centre interaction using a 'pushback' procedure. *Statistics in Medicine* 12, 1033-1045.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 41, 361-372.

Peto, R. (1982). Statistical aspects of cancer trials, in Halnan, K. E. (ed.), *Treatment of Cancer*, Chapman and Hall, London, pp. 867-871.

Piantadosi, S., and Gail, M. (1993). A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine* 12, 1239-1248.

Simon, R. (1982). Patient subsets and variation in therapeutic efficacy. *British Journal of Clinical Pharmacology* 14, 473-482.

Zelterman, D. (1990). Tests for qualitative interactions. *Statistics and Probability Letters* 10, 59-63.