Testing for Qualitative Interactions Between Stages in An Adaptive Study

Robert A. Parker

Address: Truth, Ltd., 3311 Blue Ridge Court
Westlake Village, CA 91362
Email: Bob.Parker@truthltd.com
Telephone: 347-TRUTHLTD (347-878-8458)

Abstract

I consider the underlying structure for a test of qualitative interaction of a treatment when assessing heterogeneity between stages in an adaptive trial. Since decisions about the clinical utility of a drug are based on the balance of risks and benefits, a quantitative interaction in treatment efficacy across different groups could lead to qualitatively different decisions. Thus, the difference between quantitative and qualitative interactions is not a true dichotomy. I show that the standard tests for qualitative interactions (Gail and Simon, 1985 [1]; Piantadosi and Gail, 1993 [2]) are very conservative in this application. Theoretical calculations in a simpler situation confirm that the published criteria are very conservative, which may help explain why the tests are known to have very low power to detect interaction. I introduce the concept of "minimum detectable effect", which is the smallest effect that a study could identify as statistically significant. I propose that important heterogeneity between stages in an adaptive trial be identified when two criteria are met. First, at least one individual stage must be below the overall study mean by at least the minimum detectable effect. Second, using an appropriate critical value based on simulations, there must be statistically significant heterogeneity between the stages.

Keywords: quantitative interactions; subgroup analysis; heterogeneity; minimum detectable effect

1. Introduction

The term qualitative interaction was first introduced by Peto [3]. Qualitative interaction (sometimes termed a "cross-over interaction") occurs when the treatment effect (the difference between an experimental treatment and a control group) is positive (beneficial) in some subgroups and negative (harmful) in others. In contrast in a quantitative interaction the magnitude of the treatment effect differs in different subgroups, but the direction (e.g., experimental treatment superior to control) is the same across all subgroups. In the presence of a qualitative interaction, the optimal treatment decision is different in different subgroups: some subgroups would benefit from receiving the experimental treatment, and others would be harmed by receiving the experimental treatment.

Gail and Simon [1] introduced the standard test for qualitative interaction. Using data from all subgroups, the Gail-Simon test (GS) assesses whether there is a group of subgroups in which the true treatment difference is in the opposite direction from the overall significant treatment group. Piantadosi and Gail ([2]; PG) proposed a range test, which uses only the results of the smallest and largest standardized treatment effect to identify qualitative interaction. Critical values for both tests are based on asymptotic normal theory assuming an overall significant treatment effect. There have been both theoretical extensions to these tests [4,5, 6] and extensions to a range of applications [7,8,9]. Pan and Wolfe [10] extended the concept of qualitative interactions into two categories: "severe qualitative interaction" and "slight qualitative interaction" based on whether the qualitative interaction exceeds a clinically significant difference or not.

The decision that a drug is clinically useful, however, is not based on treatment efficacy alone. All drugs have potential side effects so that the decision to use a drug invariably involves balancing potential risks and benefits. [11] In fact, "safe" means that the benefits appear to outweigh the risks. [12] Thus, it would be possible for the risk-benefit decision to differ between subgroups either because of differences in toxicity among the subgroups or differences in treatment efficacy among the subgroups, even though the drug shows treatment efficacy in all subgroups. Thus a quantitative interaction in treatment effect might lead to qualitatively different decisions, suggesting that the distinction between quantitative and qualitative interaction is one of degree rather than an absolute difference.

The issue of qualitative interactions has become more important with the recent release of the EMEA discussion document on flexible designs (Committee for Medicinal Products for Human Use, [13]) which emphasizes the need to show homogeneity across stages of the study even in the executive summary: "Using an adaptive design implies that the statistical methods control the pre-specified type I error, that correct estimates and confidence intervals for the treatment effect are available, and that **methods for the assessment of homogeneity of results from different stages are pre-planned**. [my emphasis] A thorough discussion will be required to ensure that results from different stages can be justifiably combined."

There has been some recent work on this problem. Gallo and Chuang-Stein [14] discuss practical problems with assessing heterogeneity in an adaptive study and issues in the interpretation of such heterogeneity should it be found. Friede and Henderson [15] have suggested a multi-step approach to examining whether there are changes in treatment effect in an adaptive trial with 2 stages, with a focus on changes over time. Their approach begins with a standard heterogeneity test for differences between the 2 stages from the meta-analysis framework. Should a difference be found, then they would examine the results in the first stage to determine if there is a change point in the data. If a change point is found they then test to see if there is evidence for heterogeneity between the data in the first stage after the change point and the second stage to determine whether these data can be pooled.

In practice, a drug application would only be filed if there were a significant overall benefit shown in the study, either in efficacy or in safety. Thus, heterogeneity between stages around this overall significant effect could occur in two different ways. The first, would be when there is a clear treatment benefit in each stage, although these may be quantitatively different, and the risk: benefit decision in each stage would be the same. Although this situation appears unlikely, studies sometimes are continued to obtain an adequate safety database even when there is evidence of clear treatment benefit relatively early in the study. Another situation would be when there is an overall significant treatment effect driven largely by the results in some of the stages, with relatively small treatment benefit in one or more stages of the adaptive study. Importantly, in this situation the treatment effect in some stages might not outweigh the overall risk from treatment, so that within the stage the treatment would not be considered clinically useful. The Friede - Henderson approach would not distinguish between these two circumstances.

In the case when the results in different stages would be interpreted as qualitatively different, it seems appropriate to analyze these differences as a qualitative interaction. This can be done by analyzing the results in each stage after subtracting the average overall treatment effect. After centering around the overall effect found in the study, it would be the case that the interaction test would be testing for both a positive effect in some stages and a negative effect in other stages, without a significant overall effect after centering.

In this paper I consider the case of a qualitative interaction after centering for an overall study effect as described above. This test would be used at the end of a study, when there already was evidence of an overall treatment benefit. In Section 2 I introduce notation and review both the GS and PG tests. In Section 3 I present results of numerical simulations illustrating the difference between the published criterion assuming an overall significant effect and those derived from simulations when the underlying assumption is one of no overall effect after centering. These results help explain why the GS test generally has low power, as shown previously (e.g., Piantadosi and Gail, [2]). In Section 4, I suggest a heuristic basis for setting the margin of difference between stages as the minimum detectable effect for the study and discuss the power to detect such differences using a joint clinical and statistical criterion. I conclude with a brief discussion of the results. Theoretical discussion of a simpler problem in the Appendix is included to support the results in the body of the paper.

2. Notation and Review of the Commonly Used Tests for Qualitative Interactions

For concreteness, consider the case of a two-group parallel-arm randomized study with a continuous endpoint, which has *m* stages of adaptation when concluded. Each of these *m* stages can be considered a separate sub-group, and the homogeneity of the effect across stages is a regulatory concern. Let μ denote the difference between the two treatment arms over the whole study, with standard error of treatment effect estimated by σ. Let μ_i and σ_i denote these quantities for the i^{th} stage of the study, respectively, $i = 1,..., m$. We assume that in at least one stage the difference μ_i is considered minimal, raising the concern that results should be interpreted differently in different stages, i.e., that there is a qualitative difference in decisions between stages. Let $\delta_i = (\mu_i - \mu)/ \sigma_i$ be the standardized deviation from the overall study mean during the i^{th} stage, so that δ_i is the standardized deviation constrained so that the average effect before standardization by the standard error is zero. The null hypothesis would be that all δ_i are equal to zero, and the alternative hypothesis would be that there is at least one δ_i less than zero and one δ_{i^*} greater than zero, $i^* \neq i$. For convenience, I assume that the number of subjects in each stage is the same, although the simulation program used to obtain critical values in this paper does not require this assumption.

In the Gail-Simon ([1]; GS) test, homogeneity across stages is tested using the smaller of the sum of the standardized differences squared greater than zero and less than zero, $\left(\sum_{\delta_i > 0} \delta_i^2, \sum_{\delta_i < 0} \delta_i^2\right)$ respectively. The test criterion for various levels of α is given in Table 1 of their paper, based on considering the special case where the effect in the first subgroup is infinitely large, and the effect in the other subgroups is 0, and then determining the null distribution for the sum of squares for this point. Thus, in the case *m* = 2, one of the two subgroups is guaranteed to have a large effect, since the overall effect is assumed to be significant, and the question becomes does the other sub-group have a significant effect in the opposite direction. For this reason, a one-sided P-value of 0.05 is appropriate. This is equivalent to a two-sided P-value of 0.10. As a chi-square type statistic is being used this gives a test criterion of 2.71 for α=0.05. Piantadosi and Gail ([2]; PG) give a simpler approach, using only the smallest and largest standardized differences, with a test criterion of 1.64 for α=0.05 for two groups. This is equivalent to the GS test since a standardized difference, rather than a standardized difference squared is used.

3. Simulated Critical Values When the Results are Constrained

In both papers, critical values are derived assuming that one subgroup has an infinitely large value. This assumption is not true when testing for heterogeneity after removing the overall treatment effect, however. In such a situation we are focusing on whether there is evidence that the stages in the adaptive study are significantly different from the overall effect.

As discussed in the Appendix, actual critical values depend on the size of the individual stages and deriving critical values for the adaptive trial problem appears intractable.

Therefore, I have used simulation to estimate the actual test criterion and power for both the GS and the PG test. It is possible, however, to obtain theoretical results for a simpler problem for the PG test, which does not require that the results from different stages be constrained to sum to zero. Results in the Appendix, shown for the full range of stages (subgroups) included in the PG paper [2] show that the published critical values are conservative for this simpler problem in all cases, becoming increasingly conservative as the number of stages increases. Furthermore, the Appendix shows that the results from the simulation approach and the theoretical results for this simpler problem are consistent. Finally, in the Appendix I show that a little known test suggested by Azzalini and Cox in 2004 [16] seems to provide more appropriate critical values for this simpler problem than the PG approach.

I present simulation results for two cases of an adaptive trial. One case is a mega-trial (10,000 subjects per treatment arm per stage), to provide the best chance for the GS and PG test to perform well since they assume asymptotic normality. The second case is a single large total study of 1500 patients, 750 in each of two arms. This is presented as an illustration only; software is available from the author to simulate the critical value for a specific problem. Such a study provides approximately 95% power to detect a 0.2 standard deviation difference for a continuous endpoint between the two arms at $\alpha=0.05$, two-sided. This design characteristic will be used later in Section 4 when developing a heuristic criterion for substantial heterogeneity between stages.

In Table 1, results are presented for both the GS and PG test for *m* stages of equal size, *m* ranging from 2 to 30. Although few adaptive trials will have even 5 stages, results are presented for up to 30 stages (subsets) as in the original publications [1; 2] so that results from their theoretical calculations can be compared to the simulation results in this paper. For each *m*, I assumed equal size study groups (n_m), either 10,000 for the "mega" study or the largest integer less than or equal to 750/*m* subjects in each treatment arm for the "practical" study. For the practical study, the total number of subjects ranged from 1,472 to 1,500. The following procedure was used for simulations. First, treatment effects estimates were generated for each of the *m* stages as the difference between two random observations from an N(0,1/ n_m) distribution. The average of the *m* effects in the simulation was subtracted from each of the original effect so that the sum over the *m* stages was zero, i.e., results are constrained to sum to zero. This precisely mimics the calculations of differences from the overall effect at each stage in an adaptive study, with the simulated results having the same relationship between stages that would be induced between stages in an adaptive study by following the proposed procedure. The variance of the mean difference in each stage was calculated as the sum of two random observations from a $\chi^2_{n_m-1}$ distribution divided by (n_m-1) n_m. The GS and PG statistics were calculated from each simulated set of results. Simulations were done 1,000,000 times for m $\leq$ 10 and 500,000 times for m > 10. The Type I error is the percent of cases in which the test statistic equals or exceeds the published test criterion, and the critical values in Table 1 are the 95th percentile of the simulated test statistics.

<TABLE 1 HERE>

Note that even for *m* = 2 stages, the actual Type I error is no more than 2% despite a nominal α = 0.05 for both the GS and PG test. For more than 2 stages, the Type I error for both tests is well under 1%, dropping to less than 0.3% with a large number of stages. This shows that the use of the standard GS and PG test criterion for this application would be quite conservative. The critical values based on simulation for the mega study are generally smaller than for the practical study, and this difference becomes more pronounced as the number of stages increases.

4. Criteria for Identifying Significant Heterogeneity Between Stages

4.1 Type I Error for Possible Criterion for Significant Evidence of Heterogeneity

Given that we are concerned that the treatment effect in at least one stage is sufficiently small that the results would be interpreted as qualitatively different between stages, a test for qualitative difference would seem to be more appropriate than a test for quantitative interaction. For convenience, assume that a treatment benefit is coded as positive, so that we are looking for a stage in which the treatment effect is substantially less than the overall average effect. Adopting the idea of a "severe qualitative interaction" from the Pan and Wolfe [10] paper, it would seem reasonable to require that the difference be important. As mentioned above, the sample size used in the illustration provides 95% power for a difference of 0.2 standard deviations in the total study. Thus, one could argue that unless the treatment effect during one of the adaptive stages was at least 0.2 population standard deviations below the overall effect in the study, then there would not be a meaningful difference between the different stages, based on the sponsor's criterion for an important effect. As δ_i is the standardized difference between the treatment effect in the *i*th stage and the overall study mean, this criterion implies that $\min_i (\delta_i) < -0.2$. In Table 2, results are given for up to *m* = 10 stages, although it is likely that most adaptive studies would have far fewer stages. Results are presented for this extremely large number of stages so that the properties of the proposed approach can be better understood. The first two columns of results are for the case when $\min_i (\delta_i) < -0.2$. When no requirement for statistical significance is imposed (column 2), the proportion of false positives increases rapidly, and is over 5% when *m* = 4, and is over 10% for *m* = 5, since no adjustment is being made for selecting the most extreme value. Requiring statistical significance using the empirical significance criterion from the simulations protects against this problem. In the cases of most interest, however, with two or three stages, the Type I error is substantially below the nominal 0.05 level whether or not statistical significance is required. If statistical significance is also required, than the Type I error is still below the nominal 0.05 level even with *m* = 5 stages.

<TABLE 2 HERE>

As an alternative, the smallest treatment effect that would be detected as statistically significant in the overall trial could be used. I term this the minimum detectable effect. There are two advantages to such a criterion. First, this criterion ensures that if the overall study was statistically significant that the effect in the worst stage would at least be in the same direction as the overall study. Second, this criterion is unaffected by the power criterion used when designing a study. Thus, it would not be affected by whether the study were designed for 80% power with a treatment difference of 0.15SD or for

95% power with a treatment difference of 0.20SD. For a 1500 person study, the minimum detectable effect for the example is 0.102 SD, which would be detected with 50% power. I use a slightly smaller criterion ($\min_i (\delta_i) < -0.1$) to ensure that the treatment effect is each stage is at least slightly positive. The rightmost two columns of Table 2 show the results for $\min_i (\delta_i) < -0.1$. Using this criterion for a meaningful difference and requiring statistical significance as well seems to provide quite acceptable performance, with the nominal Type I error 0.05 as *m* varies from 2 to 10.

4.2 Power for Proposed Criterion When There Is Heterogeneity

Given a joint criterion that there that there is at least one stage more than the minimum detectable effect below the overall average and that the test for a qualitative interaction be statistically significant, the next question would be the power of this approach to detect differences when in fact there are stages which are substantially below the overall average.

To examine this, results for the mean and variance of the treatment effect at each stage were calculated. Before centering the treatment effects, however, a fixed fraction of the stage-specific standard deviation was subtracted from the simulated difference for each of the stages in the simulation with a defined difference. For the remaining stages, an amount was added so that the overall expected value would be zero across all the stages. For example, when simulating a study with *m* = 4 stages and $\delta_1 = -0.20$, I would subtract 0.20 σ_1 from the treatment effect estimated in the first stage, μ_1. I would add 0.0667 (=0.20/3) σ_i, i=2, 3, 4 to the estimated treatment effect μ_i. The` treatment effect estimates over all 4 stages were then constrained so that the overall sum of differences was zero. Results are shown in the Figure. Again, results are presented for up to 10 stages so that the properties of the proposed approach can be better understood, even though few adaptive studies would have even 5 stages.

<FIGURE HERE>

In the Figure, each of the individual curves shows how power decreases as the number of stages increases, with the number of stages with low response and the magnitude of the low response fixed. The left panel shows how power increases as the magnitude of the amount below the average increases in a single stage. The right panel shows how power increases as the number of stages with a low response increases. These results are consistent with the material in Piantodosi and Gail [2], Tables 3 and 4, which show that a larger discrepancy is needed for the same power as the number of stages increases, and that power increases as the number of stages with substantially discrepant values increase.

5. Discussion

There are several important points in this paper. First, although there is a sound theoretical justification for the significance criterion for both the GS and PG tests, this criterion assumes that there is an overall significant result for the entire study. If this requirement is not met, which would occur when looking for heterogeneity between stages in an adaptive trial after adjusting for the overall study effect, then the published significance criteria are extremely conservative and lead to very low power to detect interactions.

I have proposed an approach to assess heterogeneity in an adaptive clinical trial combining both a criterion using the minimum detectable effect of the overall study, while also requiring statistical significance for the test of heterogeneity. Importantly, the use of the minimum detectable effect ensures that there is at least a slight benefit in each stage of the study. The significance criterion is based on simulations, rather than the published PG criterion. Identifying a stage with a substantially smaller treatment benefit than the overall study is only a first step, however. Once such a stage is identified, then the risk-benefit ratio needs to be considered within that stage. Only if the risk-benefit ratio based on the actual treatment effect within the stage suggested that the treatment was not clinically useful would there be a suggestion of a qualitative interaction between stages.

This approach treats the results of different stages as independent, which is consistent with common approaches for the analysis of adaptive trials. For example both the combination of P-values approach originally proposed by Bauer and Köhne [17] and the combination of test statistics approach [18] treat results of separate stages as independent. Müller and Schäfer [19] actually describe their approach as "analogous to considering data from before and after an interim analysis point as two separate studies." [19, page 890].

The process of adjusting the results across stages for the overall study effect may induce some relationship between stages, but this has been fully incorporated in the simulation process as well, so the significance criterion would reflect such a relationship. Of possibly greater importance, the decision at each stage of an adaptive study might be "stop for efficacy", "stop for futility" or "continue the study (with or without modifications)". The "stop for futility" decision at any stage implies that the study is not significant overall, so no regulatory filing would be made and concerns about qualitative interaction between stages would not be of interest. However, at either the last stage (when the study has hit the boundary on accrual / events), or at an earlier stage when the study is stopped for efficacy, the question of whether results across stages are qualitatively different would be important. One could easily imagine, for either reason for stopping, that the results are similar across stage (with a bigger treatment effect then planned when a study is stopped early for efficacy) so that the stages are homogeneous. In this situation, the decision at earlier stages to continue would occur because the accumulated evidence, although encouraging, was not sufficiently strong to demonstrate efficacy and safety convincingly. Alternatively, there could be substantial heterogeneity between stages. Unless the results were sufficient to trigger the "stop for futility" decision at the first stage, there could well be a relatively small effect at any of the stages compared to the overall treatment effect, so that in general the potential for heterogeneity between stages needs to be considered.

As a final point, I have changed a quantitative interaction into a qualitative interaction problem. The idea that a treatment would always be beneficial if there is only a quantitative interaction, however, is not valid as the overall decision involves the balance between treatment benefit and risk. It seems reasonable that the concern about homogeneity in an adaptive trial is not primarily that there is a bigger treatment effect in one stage and a smaller but still important treatment benefit in another stage. Rather, the concern would be that the overall results reflect a strong effect in one stage, and a relatively marginal effect in another. These results would be interpreted as a qualitative difference in the decision that a drug is useful, even though there is only a quantitative interaction in treatment efficacy. Since such results would be interpreted as a qualitative difference, I believe that the test used should reflect this question. The test of the stage specific difference from the overall effect, using the minimum detectable effect for the overall study to ensure that there is at least some positive benefit in each stage, and statistical significance for the test of heterogeneity, addresses this question.

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

2. Piantadosi S, Gail MH. A comparison of the power of two tests of qualitative interaction. *Statistics in Medicine* 1993;**12**:1239-1248.

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

4. Zelterman D. On tests for qualitative interaction. *Statistics and Probability Letters* 1990;**10:** 59-63.

5. Silvapulle MJ. Tests against qualitative interaction: exact critical values and robust tests. *Biometrics* 2001;**57:**1157-1165.

6. Li J, Chan ISF. Detecting qualitative interactions in clinical trials: an extension of range test. *Journal of Biopharmaceutical Statistics* 2006;**16:** 831-841.

7. Wiens BL, Heyse JF. Testing for interaction in studies of noninferiority. *Journal of Biopharmaceutical Statistics* 2003;**13**:103-115.

8. Yan X. Test for qualitative interaction in equivalence trials when the number of centres is large. *Statistics in Medicine* 2004;**23**:711-722.

9. Chen YHJ, Liu GH Frank. Testing for crossover of two hazard functions using Gail and Simon's method. *Journal of Biopharmaceutical Statistics*, 2006;**16**:313-326.

10. Pan G, Wolfe DA. Test for qualitative interaction of clinical significance. *Statistics in Medicine*, 1997;**16**:1645-1652.

11. Juhn P, Phillips A, Buto K. Balancing modern medical benefits and risks. *Health Affairs*, 2007;**26**:647-652.

12. Food and Drug Administration (2002). The FDA's drug review process: ensuring drugs are safe and effective. *FDA Consumer Magazine*, 36(4). Available online with revisions at http://www.fda.gov/Drugs/ResourcesForYou/Consumers/ucm143534.htm, downloaded on July 28, 2009.

13. Committee for Medicinal Products for Human Use (2007). *Reflection paper on methodological issues in confirmatory clinical trials planned with an adaptive design.* Document Reference: CHMP/EWP/2459/02. European Medicines Agency: London

14. Gallo P, Chuang-Stein C. What should be the role of homogeneity testing in adaptive trials? *Pharmaceutical Statistics* 2009;**8:**1-4. DOI:10.1002/pst.342

15. Friede T, Henderson R. Exploring changes in treatment effects across design stages in adaptive trials. *Pharmaceutical Statistics* 2009;**8:**62-72.
DOI: 10.1002/pst.332

16. Azzalini A, Cox DR. Two new tests associated with analysis of variance. *Journal of the Royal Statistical Society: Series B* 1984;**46:** 335-343.

17. Bauer P, Köhne K. Evaluation of experiments with adaptive interim analyses. *Biometrics* 1994;**23:**3-15.

18. Proschan MA, Hunsberger SA. Designed extension of studies based on conditional power. *Biometrics* 1995;**51**:1315-1324.

19. Müller H-H, Schäfer H. Adaptive group sequential designs for clinical trials: combining the advantages of adaptive and of classical group sequential approaches. *Biometrics* 2001;**57**:886-891.

Table 1. Simulated Results for the Gail-Simon and Piantadosi-Gail Test for Identifying Heterogeneity Between Stages in an Adaptive Study (α=0.05)

Number of Stages	Gail-Simon Test				Piantadosi-Gail Test			
	Published Critical Value[a]	Type I Error[b]	Critical Values from Simulations		Published Critical Value[d]	Type I Error[b]	Critical Values from Simulations	
			Mega Study[c]	Practical Study[c]			Mega Study[c]	Practical Study[c]
2	2.71	1.92	1.90	1.87	1.64	1.97	1.38	1.37
3	4.23	0.53	2.36	2.37	1.95	0.63	1.45	1.45
4	5.43	0.52	3.22	3.21	2.12	0.47	1.58	1.58
5	6.50	0.41	3.92	3.94	2.23	0.42	1.68	1.69
6	7.48	0.39	4.65	4.67	2.32	0.37	1.76	1.77
7	8.41	0.36	5.34	5.38	2.39	0.34	1.83	1.84
8	9.29	0.35	6.02	6.08	2.44	0.33	1.88	1.90
9	10.15	0.34	6.69	6.75	2.49	0.32	1.93	1.95
10	10.99	0.33	7.34	7.43	2.53	0.32	1.98	1.99
12	12.60	0.31	8.64	8.76	2.60	0.31	2.05	2.07
14	14.15	0.30	9.91	10.09	2.66	0.29	2.11	2.14
16	15.66	0.29	11.15	11.38	2.71	0.27	2.16	2.20
18	17.13	0.28	12.39	12.67	2.75	0.27	2.21	2.25
20	18.57	0.29	13.63	13.96	2.78	0.28	2.25	2.30
25	22.09	0.27	16.62	17.16	2.86	0.27	2.33	2.39
30	25.50	0.26	19.58	20.37	2.92	0.27	2.40	2.48

[a] Gail and Simon [1], Table 1, column 4, significance level 0.05

[b] Percent of simulations of the mega study statistically significant using published significance criteria.

[c] Mega study has 10,000 subjects / treatment arm / stage, i.e., 40,000 total subjects for m=2 stages to 600,000 total subjects for m=30 stages, to illustrate asymptotic results. Practical study has 1,472-1,500 subjects total.

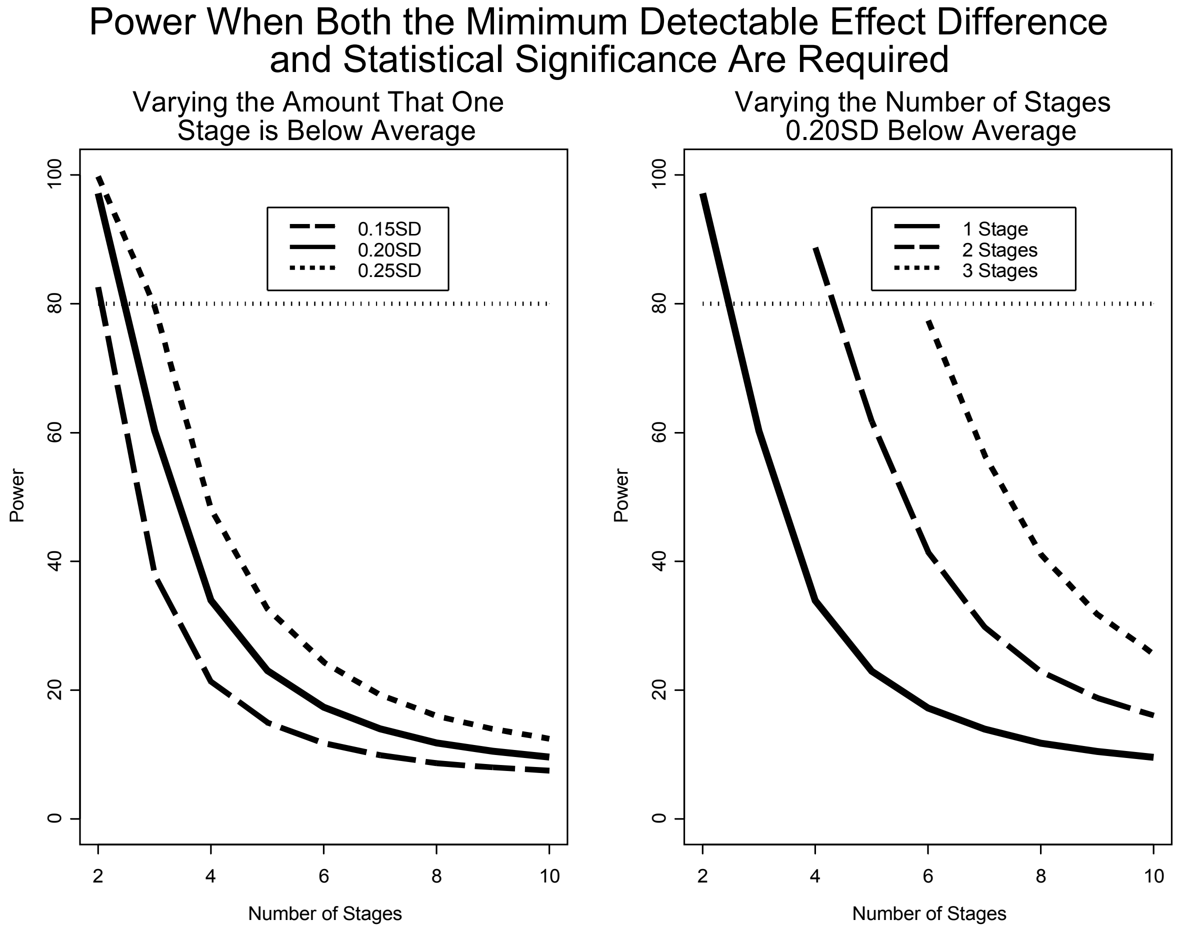
[d] Piantadosi and Gail [2], Table 1, column 4, significance level 0.05

Table 2. Type I Error of Detecting a Stage Substantially Below the Average Effect

Number of Stages (*m*)	Percent of simulations in which the smallest treatment effect is more than 0.2 SD below the overall mean		Percent of simulations in which the smallest treatment effect is more than 0.1 SD below the overall mean	
	No requirement for statistical significance	Statistical significance using simulation-based criterion also required	No requirement for statistical significance	Statistical significance using simulation-based criterion also required
2	0.01	0.01	5.34	5.00
3	0.96	0.73	25.66	5.03
4	5.27	2.36	49.69	5.05
5	13.37	4.59	69.72	5.05
6	24.46	5.01	83.42	5.01
7	37.26	4.94	91.61	4.94
8	50.44	4.94	96.08	4.94
9	62.08	5.01	98.27	5.01
10	72.00	5.02	99.26	5.02

Figure Legend

Power to detect heterogeneity between stages in an adaptive trial when both the minimum detectable effect and a statistically significant qualitative interaction are required. The minimum detectable effect for the overall study is approximately 0.10 SD. Statistical significance was determined using the Piantadosi-Gail test with test criteria based on simulations rather than the published criteria. The left panel shows the power as the total number of stages in the study varies when one stage in the adaptive trial has a true treatment effect less than the overall average by 0.15, 0.20, or 0.25 of the stage specific SD. Power decreases as the total number of stages increases, and increases as the magnitude of the difference increases. The right panel shows power as the total number of stages in the study varies as the number of stages with a true treatment effect less than the overall average by 0.20 of the stage specific SD increases from 1 to 3. Power increases as the number of stages with a low treatment effect increases.



Power When Both the Mimimum Detectable Effect Difference and Statistical Significance Are Required

Varying the Amount That One Stage is Below Average

Varying the Number of Stages 0.20SD Below Average

Appendix. Theoretical Considerations in a Simpler Problem

I am not able to present a theoretical justification for the results presented in the body of the paper. Such results would require solving for the test criterion over the entire distribution of results over all groups, as the adjustment for the mean effect varies as the result in each stage varies. In this Appendix, I present results for a simpler problem. As in the body of the paper, no overall treatment effect is assumed, but unlike the rest of the paper there is no adjustment such that the overall treatment effect across stages is zero. In this Appendix I show (a) that the Piantadosi-Gail ([2]; PG) test is even more conservative than shown in the body of the paper; (b) that test criterion need to be based on t-distributions, incorporating the actual group sizes, rather than on an assumption of asymptotic normality; and (c) that a test previously proposed by Azzalini and Cox [16] is much more appropriate for this simpler problem than the more widely known PG test. Given the close relationship of the PG test to the standard Gail-Simon ([1]; GS) test, these results would also be expected to hold for the GS test as well. Simulations (not shown) confirm this.

A.1 Theoretical Critical Values

For the PG test when no adjustment is made for the overall mean μ, it is possible to derive the critical values from theoretical considerations. Let λ_m be the critical value for the analysis with *m* stages. We know that (a) the smallest value of μ_i / σ_i must be less than $-\lambda_m$; (b) the largest value of μ_i / σ_i must be greater than λ_m; and (c) the values of the *m-2* other groups are between the lowest and the highest value. As the PG (and GS) derivations use asymptotic normality, let $\phi(x)$ be the normal density function and $\Phi(x)$ be the cumulative normal density function from $-\infty$ to x. The critical value for a significant difference at level α can be obtained using numerical integration from

$$\alpha = m\,(m-1)\int_{-\infty}^{-\lambda_m}\varphi(x)\left[\int_{\lambda_m}^{\infty}\varphi(y)\left(\Phi(y)-\Phi(x)\right)^{m-2}\,dy\right]dx \qquad \text{(A.1)}$$

This calculation assumes that the ratio μ_i / σ_i is normally distributed. As the mean and standard deviation are both estimated from the same set of data, however, the ratio μ_i / σ_i actually follows a t-distribution, however, Calculating a critical value accounting for varying group sizes would involve summation of the integrals for the *m(m-1)* individual combinations of lowest and highest groups. As a simplification, one could assume that all groups are the same size, as done in the body of the paper. Let *t(x,n)* denote the probability density function for the t-distribution at *x* with *n* degrees of freedom and T(x,n) the cumulative density function from $-\infty$ to x. Then the critical value for a significant difference at level α can be obtained using numerical integration from

$$\alpha = m\,(m-1)\int_{-\infty}^{-\lambda_m}t(x,n)\left[\int_{\lambda_m}^{\infty}t(y,n)\left(T(y,n)-T(x,n)\right)^{m-2}\,dy\right]dx$$

(A.2)

When the number of groups is small and the individual group sizes are large, there would only be a small difference between the values calculated for λ_m from (A.1) and (A.2). The critical value obtained from (A.1) would in general be a lower limit for the true critical value for an actual problem with finite data, since the term $\left(\Phi(y)-\Phi(x)\right)^{m-2}$ would be larger than $\left(T(y,n)-T(x,n)\right)^{m-2}$. However, in cases where the number of subgroups is large, which in practice implies that the size of each subgroup is relatively small, these differences can be substantial. For example with m = 30 and n = 50 (25 in each treatment group), for a relatively large total study with 1500 total patients, the critical value assuming asymptotic normality from equation (A.1) is 2.39 while the critical value using the t-distribution in equation (A.2) is 2.47. Importantly, both values are substantially smaller than the published critical value for the PG test of 2.92.

<Table A.1 here>

Simulations using the approach outlined in Section 3 of the paper, without the step of centering the results, show consistency between the theoretical criterion from (A.2) and the results from simulations, with differences between the theoretical result and the simulation based test criteria of 0.01 after rounding in three cases. In addition these simulation results show that the PG test is even more conservative than shown in the body of the paper. For example, for the case of two subsets the chance of a Type I error when the true treatment effect in all groups is zero would be 0.005 (2 x 0.05 x 0.05), which is only one-tenth of the nominal value, α=0.05, compared to a Type I error of 0.0197 when results are adjusted (Table 1). As shown in the fifth column of Table A.1, the result from the simulation confirms this.

A.2 Alternative Approach to Testing for Qualitative Interaction

Shortly before the GS paper was published, an alternative approach was presented by Azzalini and Cox [16], which, although discussed by Gail-Simon [1], appears to be much less known. In their approach, the problem is formulated both in terms of a variable number of treatment groups (m_1 in their notation; assumed 2 in this paper) and a variable number of subgroups (m_2 in their notation, denoted m in this paper). They defined a qualitative interaction as when there existed at least one pair of treatments for which there existed two subgroups such that the treatment difference was positive in one of the subgroups and negative in the other above a certain significance criteria. When considered in the context of m_1 = 2 treatment groups, this is the same approach as used in the PG test.

Formula (9) in the Azzalini and Cox paper [16] gives the significance criteria as

$-\Phi^{-1}\left[\left\{-\frac{2\log(1-\alpha)}{m_1(m_1-1)m(m-1)}\right\}\right]$ where Φ is the standard normal distribution. It is worthwhile to understand the logic underlying this criterion. One is adjusting the overall significance criterion for the test criterion for the selection of the two specific subgroups with extreme results [*m(m-1)* possible combinations of subgroups] for each of the $m_1(m_1\text{-}1)$ possible treatment combinations. For example, with *m* = 2 groups and 2 treatments, the probability of exceeding the test criteria (0.99) for each individual comparison is 0.161. Allowing for the two possible orders for the test (e.g. treatment 1 > treatment 2 in the first subgroup or in the second subgroup), then the overall probability of a significant result is 2 x 0.161 x 0.161 = 0.052. The results are similar for other values of *m* as well. Thus, this test would seem to reflect an attempt to identify qualitative interaction in the situation when there was no overall significant treatment effect and without any attempt to center the results as in the adaptive situation.

As such, results from this approach can also be usefully compared to the simulation results in Table A.1. Azzalini and Cox [16] give results for m_1 = 2 treatment groups for m = 3, 4, and 6 subgroups in their Table 1. For all cases, the criterion presented in the Azzalini-Cox paper, theoretical results, and the results from the simulations are close. The results from my simulations go from slightly under the published criteria when m = 3 (by 0.02) to slightly over (by 0.05) when m = 30. For m = 2, the difference is only 0.01, even though Azzalini and Cox did not give an explicit result in their table 1. All these three results are very much smaller than the published PG criteria.

Table A.1. Results from Theoretical Calculations and Simulations for the Piantadosi-Gail Test Applied to the Situation When There is No Overall Significant Treatment Effect

Number of stages	Test Criteria for the Piantadosi-Gail Test (α = 0.05)			Results from Simulation		
	Published Criteria[a]	Theoretical Results[b]	Azzalini-Cox[c]	Type I Error[d]	Critical Values	
					Mega Study[e]	Practical Study[e]
2	1.64	1.00	0.99	0.50	1.00	1.00
3	1.95	1.31	1.33	0.38	1.30	1.31
4	2.12	1.48	1.51	0.34	1.48	1.48
5	2.23	1.61	1.64	0.32	1.60	1.61
6	2.32	1.71	1.74	0.30	1.70	1.70
7	2.39	1.78	1.81	0.28	1.77	1.79
8	2.44	1.85	1.88	0.28	1.84	1.85
9	2.49	1.91	1.93	0.28	1.89	1.91
10	2.53	1.96	1.98	0.29	1.94	1.96
12	2.60	2.04	2.06	0.27	2.02	2.04
14	2.66	2.11	2.13	0.27	2.09	2.11
16	2.71	2.17	2.18	0.26	2.14	2.18
18	2.75	2.23	2.23	0.26	2.19	2.23
20	2.78	2.28	2.27	0.26	2.23	2.28
25	2.86	2.38	2.36	0.25	2.32	2.38
30	2.92	2.47	2.42	0.26	2.39	2.47

[a] Piantadosi and Gail [2], Table 1, column 4, significance level 0.05

[b] Equation A.2 for practical study, 1,488-1500 subjects total.

[c] Authors calculation for all cases except m = 3, 4, 6 which are abstracted from Azzalini and Cox [16], Table 1, m_1 = 2, page 338. Note that Azzalini and Cox explicitly do not provide a result for m_1 = 2 treatments and m_2 = 2 subgroups in their paper.

[d] Percent of simulations of the mega study statistically significant using published significance criteria.

[e] Mega study has 10,000 subjects / treatment arm / stage, i.e., 40,000 total subjects for *m*=2 stages to 600,000 total subjects for *m*=30 stages, to illustrate asymptotic results. Practical study has 1,472-1,500 subjects total.