Detecting Qualitative Interaction: a Bayesian Approach

Emine Özgür Bayman[1,2], Kathryn Chaloner [2,3], Mary Kathryn Cowles [3,2]

[1] *Department of Anesthesia,* [2] *Department of Biostatistics,* [3] *Department of Statistics,*
The University of Iowa, Iowa City, IA

SUMMARY

Differences between treatment effects between centers in a multi-center trial may be important. These differences represent treatment by subgroup interaction. Peto (1982) defines qualitative and quantitative interaction. Qualitative interaction occurs when the simple treatment effect in at least one subgroup has a different sign than in other subgroups: this interaction is important and is important to detect. Interaction which is not qualitative is called quantitative: this is common and is often not important. In this paper, based on two motivating multi-center clinical trials with binary responses, a hierarchical model is used, with exchangeable mean responses between subgroups to each treatment; this reflects that the treatment responses are similar but different. The posterior probability of qualitative interaction is calculated and the Bayes Factor, which compares this posterior probability to the corresponding prior probability, is proposed as a diagnostic for qualitative interaction. This Bayes Factor can be used to derive a significance test which can be compared to other frequentist tests for qualitative interaction. The frequentist power of the Bayesian test is examined and compared to two other commonly used approaches due to Gail and Simon (1985) and Piantadosi and Gail (1993) for different fixed combinations of treatment effects. The impact on power of imbalance between the sample sizes in each subgroup is examined and the test based on the Bayes Factor typically has better power for unbalanced designs, especially for small sample sizes. An exact test based on each of the three test statistics is also suggested and applied to two situations assuming the random effects model.

Keywords: Bayesian power, multi-center clinical trials, qualitative interaction, subgroup power.

1. INTRODUCTION

Medical researchers typically examine response to treatment for different types of subjects in a clinical trial. For example, they may want to know whether a treatment effect is different in older subjects versus younger subjects, or in men versus women. If, for a group of subjects, a treatment is not beneficial or even is harmful, but it is beneficial for others, this should be discovered. Examining only the overall treatment effect may obscure an important effect of treatment in particular subgroups [1].

A *subgrouping* is a partition of patients into mutually exclusive subsets or subgroups based

*Correspondence to: Emine Özgür BAYMAN, Department of Anesthesia, The University of Iowa, Iowa City, IA
†E-mail: emine-bayman@uiowa.edu

on values of one or more variables [2]. *Subgroup analysis* is analyzing the treatment effect in each subgroup category.

Comparing two treatments for each subgroup category separately increases the probability of at least one type I error [3]. If there are many categories, control of error rates with traditional approaches is difficult. In addition, the power of the test for interaction of a particular magnitude is lower than that of the test for an overall effect of the same magnitude. Therefore, applying separate subgroup analysis after showing statistically significant treatment-by-subgroup interaction may not be an effective approach [3].

Follmann provides a survey of statistical techniques for subgroup analysis and interaction [6]. He also calculates the chance of having a p-value of less than 0.05 in one of two subgroups, when the overall p-value is not significant. This probability is surprisingly large.

Peto [7] defines *qualitative* interaction (QI) as arising when the sign of the true treatment differences varies between subgroups. When the magnitude of the treatment benefit varies, but the sign does not, then there is *quantitative* interaction. Quantitative interactions are unimportant and common while qualitative interactions are important and less likely [7]. A qualitative interaction implies that the subgroup populations should be treated differently and the overall result does not apply to all subgroups.

1.1. Review of Frequentist Methods

The standard test of interaction does not distinguish between qualitative and quantitative interaction. Gail and Simon (G&S) [8] proposed a likelihood ratio test (LRT) constructed to detect qualitative interaction. To apply this test to mutually exclusive subgroups, the estimates of treatment effects in all subgroups are assumed to be normally or approximately normally distributed with a known, or accurate estimate of, variance. This test is widely used to detect qualitative interaction.

Silvapulle [9] stated that the G&S test is likely to be sensitive to outliers, and that the asymptotic critical values given by G&S might not be appropriate in small samples. He suggested replacing the least squares estimator by a robust estimator, the M-estimator, and derived the exact critical values. Russek-Cohen and Simon [10] generalized the procedure of G&S [8] to the case of two correlated estimates of treatment effect.

Piantadosi and Gail (P&G) [11] proposed an alternative test of qualitative interaction based on the range test, which involves checking the minimum and the maximum observed treatment differences over subgroups. G&S and P&G both use LRT to test for any interaction and then to use the test of qualitative interaction only if the LRT for any interaction is significant. When there are only two subgroups, P&G's test of qualitative interaction is equivalent to G&S's test. P&G [11] claims that the range test should be more powerful when a treatment effect is harmful in only a few subsets and beneficial in most of the subsets or vice versa, whereas the G&S test [8] should be more powerful when the new treatment is harmful in several subsets and beneficial in several other subsets. This claim is examined in Section 3 for examples of binary responses.

Li and Chan [12] extended P&G's range test to use all observed treatment differences rather than only the minimum and the maximum. Pan and Wolfe [13] developed a test of qualitative interaction of clinical significance which tests whether the difference in treatments is bigger than a pre-defined value. Yan and Su [14] presented a review of testing for qualitative interactions. Ciminera et al. [15] proposed a different ("push-back") procedure to test for

qualitative treatment-by-center interaction. Boos et al. [16] also proposed a new measure called the Interaction Magnitude. Wiens and Heyse [17] compared several tests of interaction for non-inferiority studies.

All of these authors recommend first examining the usual treatment-by-subgroup interaction, and if the interaction term is not significant, basing the treatment recommendation on the overall result combining all subgroups. They also recommend that the number of subgroups be small. In addition, if the overall treatment effect of the study is not significant, drawing definitive conclusions from subgroup analyses should be avoided [6]. In this paper, our Bayesian approach will be compared to the G&S and P&G tests of qualitative interaction.

1.2. Review of Bayesian Tests of Interaction

With the Bayesian approach, the multiple comparison issue is different [18, 19]. There is no need to adjust for experimental error rate with the Bayesian approach.

In their 1991 paper Dixon and Simon (D&S) [19] proposed a new Bayesian method for the analysis of subgroups in clinical trials with binary covariates. They used vague prior distributions for all of the regression coefficients except the treatment-by-covariate interactions, which were assumed to be exchangeable and from a Normal distribution with mean zero. However they commented that, if any subset is a priori of special interest, that subgroup should be excluded from the interaction vector, and the exchangeability assumption should not be applied to that subgroup. Because of the exchangeability of interaction terms, shrinkage occurs in the Bayesian posterior point and interval estimates. Their method therefore incorporates the prior belief that qualitative interactions are not likely to happen.

In a later paper, Simon [3] proposed a Bayesian approach for subgroup analysis under the assumption of no qualitative interactions, and used informative prior distributions for interaction effects. White et al. [20] used a questionnaire to elicit expert beliefs regarding treatment and covariate effects in clinical trials. Similar to Simon [3] they use non-informative prior distributions for main effects of treatment and prior distributions based on expert beliefs for interaction terms.

The new approach developed here incorporates exchangeable mean responses to treatment between subgroup categories. This represents a prior belief that responses in different subgroup categories will be similar, as well as a belief that qualitative interaction is unlikely.

2. The MODEL and DEFINITIONS

The treatment effect in subgroup j is represented by ϕ_j, $j = 1, ..., N$. Since qualitative interaction occurs when at least one subgroup's treatment effect is in the opposite direction to that in other subgroups, to test for a qualitative interaction the hypothesis to be tested is:

$H_0 : \phi_j > 0$ for all j or $\phi_j < 0$ for all j,
H_1 : There exists at least one pair $j \neq j'$ such that $\phi_j > 0$ and $\phi_{j'} < 0$.

Define θ_{ij} to be either the mean or a function of the mean of the response y_{ijk}, as in a generalized linear model (GLM), of subject k assigned to treatment i in subgroup j. Then $\phi_j = \theta_{2j} - \theta_{1j}$. If $\phi_j > 0$, the new treatment is superior to the standard therapy in subgroup

j. Assume that

$$\phi_j \mid \mu, \omega^2 \overset{ind}{\sim} \text{Normal}(\mu, \omega^2). \tag{1}$$

This represents the treatment effects, ϕ_j, being exchangeable between subgroups and being a sample from a normal distribution with mean μ and standard deviation ω.

A prior distribution for μ and ω is required. The prior probability of qualitative interaction, $Pr(QI)$, is as follows:

$$\begin{aligned} Pr(QI) &= 1 - [Pr(\phi_1 > 0, ..., \phi_N > 0) + Pr(\phi_1 < 0, ..., \phi_N < 0)] \\ &= 1 - \int\int [\prod_{j=1}^{N} Pr(\phi_j \mid \mu, \omega) > 0 + \prod_{j=1}^{N} Pr(\phi_j \mid \mu, \omega) < 0] dp(\mu, \omega) \\ &= \int\int \{1 - [\Phi(\frac{\mu}{\omega})]^N - [1 - \Phi(\frac{\mu}{\omega})]^N\} dp(\mu, \omega) \end{aligned} \tag{2}$$

After specification of a prior distribution for μ and ω, the double integral in Equation 2 may be approximated by Monte Carlo integration. The posterior probability of qualitative interaction based on data $\underline{y}$, $Pr(QI \mid \underline{y})$, will also be computed numerically, typically by Markov Chain Monte Carlo.

Our Bayesian test quantifies the evidence for qualitative interaction through the Bayes Factor, BF_{01}, which can be used to conduct a hypothesis test. The BF_{01} is the ratio of the posterior odds in favor of the null to the prior odds of the null. This terminology comes from Good [21](page 36), and he attributed the method to Turing and Jeffreys [22].

$$BF_{01} = \frac{[1 - P(QI \mid y)]P(QI)}{P(QI \mid y)[1 - P(QI)]} \tag{3}$$

Table I. Interpretation of Bayes Factor (Jeffreys)

Bayes Factor	**Strength of Evidence in favor of H_1**
$BF_{01} > 1$	Null hypothesis supported
$10^{-1/2} < BF_{01} < 1$	Not worth more than a bare mention
$10^{-1} < BF_{01} < 10^{-1/2}$	Substantial evidence
$10^{-3/2} < BF_{01} < 10^{-1}$	Strong evidence
$10^{-2} < BF_{01} < 10^{-3/2}$	Very strong evidence
$BF_{01} < 10^{-2}$	Decisive evidence

Jeffreys [23](page 432) provided a scale for interpreting the value of the BF_{01} as given in Table I. Jeffreys suggested interpreting BF_{01} in half-units on the $\log_{10}$ scale [22]. Thus $BF_{01} < 10^{-1}, < 10^{-3/2}$ and $< 10^{-2}$ are interpreted respectively as *strong, very strong* and *decisive* evidence against the null hypothesis with Jeffreys' scale. Kass and Raftery [22] suggested using the more conservative version of Jeffreys' interpretation, see Table II. They suggested twice the natural log of BF_{01} which is on the same scale as the deviance and LRT statistics.

Table II. Interpretation of Bayes Factor (Kass and Raftery)

Bayes Factor	**Strength of Evidence in favor of H_1**
$BF_{01} > 1$	Null hypothesis supported
$1/3 < BF_{01} < 1$	Not worth more than a bare mention
$1/20 < BF_{01} < 1/3$	Positive evidence
$1/150 < BF_{01} < 1/20$	Strong evidence
$BF_{01} < 1/150$	Very strong evidence

A new Bayesian test is proposed in which the null hypothesis of no qualitative interaction is rejected when BF_{01} is less than some bound m.

Alternatively a test for clinically meaningful qualitative interaction can be defined similarly, for example, by calculating the Bayes Factor for

$H_0 : c_b \leq \phi_j \leq c_h$ for all $j = 1, \ldots, N$,

H_1 : There exist at least one pair $j \neq j'$ such that $\phi_j > c_b$ and $\phi'_j < c_h$.

The values c_b and c_h are chosen to represent clinically meaningful differences: benefit and harm respectively. Different null and alternative hypotheses could also be examined.

2.1. Motivating Example with Binary Responses

This method was motivated by the design of a multi-center trial for cell transplantation. The cells are processed locally at each center and variability between centers was a concern. In addition, the processed cells were to be considered for licensing at the completion of the study, and each center was to apply for a separate license. The trial was originally designed as a randomized multi-center trial with 7 centers and subjects randomized to either a transplant or usual care. The response was a binary outcome at 12 months after randomization. The sample size of 65 subjects per treatment assignment was chosen based on frequentist power calculations assuming no between center variability. A Bayesian subgroup analysis was proposed to examine the treatment effect in each center separately using a model in which treatment effects were exchangeable between centers. This study is used as a framework to study the properties of procedures for detecting qualitative interaction.

Let y_{ij} denote the observed number of successes out of n_{ij} subjects in the i^{th} treatment group ($i = 1, 2$) in subgroup j (center j, $j = 1, \ldots, N$). Also let p_{ij} denote the true underlying success probability for treatment i, center j. It is assumed that given the success probability, p_{ij}, the observed number of successes for each center, y_{ij}, is a draw from a Binomial distribution.

$$y_{ij} \mid p_{ij} \sim \text{Bin}(n_{ij}, p_{ij}). \tag{4}$$

For the logit link, $\theta_{ij} = \text{logit}(p_{ij}) = \log(p_{ij}/(1-p_{ij}))$, and assume that, the θ_{ij}'s are draws from a Normal distribution with mean μ_i and common variance σ^2 in each treatment independently for $i = 1, 2$. That is,

$$\theta_{ij} \mid \mu_i, \sigma^2 \sim \text{Normal}(\mu_i, \sigma^2). \tag{5}$$

θ_{ij} represents the log odds for treatment i at center j and ϕ_j represents the log odds ratio for center j. Treatment effects ϕ_j are

$$\phi_j \mid \mu_1, \mu_2, \sigma^2 \sim \text{Normal}(\mu_2 - \mu_1, 2\sigma^2) \tag{6}$$

and therefore define $\mu = \mu_2 - \mu_1$ and $\omega = \sigma\sqrt{2}$ in equation 1.

The prior distribution was constructed with the aid of existing data from the 7 centers. The between center variance σ^2 has an inverse gamma distribution with parameters $\alpha = 2$ and $\beta = 1.5$, so that the mean of σ^{-2} is $\frac{\alpha}{\beta} = \frac{4}{3}$. The population parameters (σ^2, μ_1, μ_2), are assumed independent and both μ_1 and μ_1 have a uniform distribution on (-4.595, 4.595). The range for the uniform distribution was chosen so that the prior conditional distribution of θ_{ij} given μ_i had a specified range. This prior distribution is reasonably uninformative, and this choice was examined in a sensitivity analysis and found to be reasonably robust.

2.2. IHAST Example

IHAST is a multi-center, prospective, randomized, partially blinded clinical trial, designed to determine whether mild intraoperative hypothermia results in improved neurologic outcome in patients with an acute subarachnoid hemorrhage (SAH) undergoing an open craniotomy to clip their aneurysms [24]. The outcome is binary: a favorable outcome is defined as Glasgow Outcome Score (GOS) equal to 1 ninety days after surgery. A subject with GOS score of 1 has "a capacity to resume normal occupational and social activities with minor physical or mental deficits or symptoms" [24]. A total of 1001 subjects were followed postoperatively and GOS evaluated on or about ninety days after surgery. Randomized treatment assignment was stratified by center (30 centers) and by time from SAH to surgery (0 to 7 days or 8 to 14 days). The primary result of the study was that intraoperative hypothermia did not improve the neurological outcome after craniotomy among good-grade patients with aneurysmal SAH (66% favorable outcome on hypothermia vs 63 % favorable outcome on normothermia, odds ratio = 1.14, 95% confidence interval: 0.88 to 1.48).

Although the overall result of the trial was not significant at the 0.05 level, the interaction between gender and treatment was significant. In the hypothermia group, 69% of males (120/174) were classified as having a good outcome, as compared to 57% (97/171) in the normothermia group. Among women, 64% (209/325) in the hypothermia group had a GOS score of 1, as compared to 66% (217/330) in the normothermia group (Table III).

Similarly, as shown in Table III, the interaction between time to surgery and treatment was also significant [24]. Among the subgroup with time from SAH to surgery of 0 to 7 days there was little difference: 64% (289/452) vs 63% (287/455) for hypothermia vs normothermia; in the subgroup with 8 to 14 days after SAH, the comparison is 83% (39 of 47) vs 61% (28 of 46) for the hypothemia to normothermia comparison. These results suggest that perhaps males benefit from intraoperative hypothermia during surgery, whereas females do not and those getting surgery 8 to 14 days after SAH benefit and other do not.

First consider gender. The null hypothesis of interaction of any kind is rejected ($p = 0.03$ by the likelihood ratio test), and so the G&S and P&G tests are applied. Neither of these tests detect qualitative difference between genders ($p > 0.20$). The prior distribution of (σ^2, μ_1, μ_2) is as in Section 2.1. To apply the Bayesian procedure, the prior probability of qualitative interaction from Equation 2 is 0.160 and the posterior probability of qualitative interaction is calculated as 0.637. The posterior probability of qualitative interaction is larger than the prior probability. The BF_{01} is 0.108, indicating 'substantial' evidence with Jeffreys' scale and 'positive' evidence with Kass and Raftery's scales. The Bayes factor suggests a possibility of

Table III. Comparison of Bayesian and frequentist tests with IHAST example

	% favor. outcome			p - values			Prior	Posterior	
	Normo.	**Hypo.**	**%diff.**	**LRT**	**G&S**	**P&G**	**Pr(QI)**	**Pr($QI \mid \underline{y}$)**	**BF**
Gender									
Male	57%	69%	12%	0.03	$p > 0.2$	$p > 0.2$	0.16	0.64	0.11
Female	66%	64%	−2%						
Time from SAH to surgery (days)									
0 - 7	63%	64%	1%	0.03	$p > 0.2$	$p > 0.2$	0.16	0.32	0.40
8 - 14	61%	83%	21%						

qualitative interaction and adds additional information over the significance test. It quantifies the evidence concisely and accurately. In addition, as described later in Section 3.5, assuming that the prior distribution generates the value of $\underline{\phi}$, an exact test of size 0.05 using the Bayes Factor gives a p-value of 0.01, and corresponding exact tests using the test statistics of G&S and P&G give p-values of 0.01 and 0.39 respectively, indicating that this interaction may be important.

Similar calculations are done for the subgroup defined by time from SAH to surgery and the data are also given in Table III. For this subgroup, the posterior probability of qualitative interaction is 0.32 giving $BF_{01} = 0.404$ which indicates 'not worth more than a bare mention' by both Jeffreys' and Kass and Raftery's scales. Even though the LRT for any interaction is significant ($p = 0.03$), neither G&S nor P&G test results are significant. These results are all consistent: there is no compelling evidence of qualitative interaction.

2.3. IHAST Example with Clinically Meaningful Qualitative Interaction

Table IV. Joint prior probabilities for ϕ_1 (treatment effect for females) and ϕ_2 (treatment effect for males) lying in different regions of $\mathbb{R}^2$ in IHAST trial

	Males		
Females	$\phi_2 < -c$	$-c < \phi_2 < c$	$\phi_2 > c$
$\phi_1 < -c$	0.38	0.04	0.04
$-c < \phi_1 < c$	0.04	0.02	0.04
$\phi_1 > c$	0.04	0.04	0.38

The qualitative interaction for gender in the IHAST trial is inconclusive and so the possibility of a clinically meaningful qualitative interaction was examined. A 10% absolute difference (65% vs 75%) in good outcomes is deemed clinically meaningful, given the additional complications from hypothermia. This corresponds to a difference of $c = 0.48$ on the log odds scale. A clinically meaningful qualitative interaction is therefore defined as occuring if $\phi_j < -c$ and $\phi_{j'} > c$ for two subgroups j and j'. Joint prior probabilities for treatment effect of females and males lying in different regions of $\mathbb{R}^2$ are given in Table IV. For example, the joint prior probability of both ϕ_1 and ϕ_2 being less than c is 0.38 which is also the joint probability

that both ϕ_1 and ϕ_2 being larger than c. This reflects the positive prior correlation making qualitative interaction unlikely. The prior probability of clinically meaningful qualitative interaction is 0.08 (0.04 + 0.04) and the posterior probability of qualitative interaction is effectively zero. Almost all of the posterior probability is concentrated on two possibilities: one where the effect of hypothermia is not clinically meaningful for either males or females (posterior probability 0.52) and the second where the effect of hypothermia is clinically meaningful for males, but not for females (posterior probability 0.48). The possibility that hypothermia is clinically harmful in one subgroup has been ruled out.

3. SIMULATION STUDIES for MOTIVATING EXAMPLE

3.1. Design of Simulations

For ease of notation, let $\underline{\phi} = (\phi_1, \ldots, \phi_N)$ denote the odds ratios of the N centers. Assume that there is equal allocation within centers: $n_{1j} = n_{2j}$ for all $j = 1, \ldots, N$. Denote $n_j = n_{1j} + n_{2j}$, $\mathbf{n} = (n_1, \ldots, n_N)$ and $n_\cdot = \sum_{j=1}^{N} n_i$. The simulation studies are based on the motivating example of Section 2.1 with $N = 7$, and binary responses. The simulation study examines the ability of BF_{01} to detect qualitative interaction, if present, compared to the power of the methods of P&G and G&S. The type I error probability is also examined. Before starting the simulations, the prior probability of qualitative interaction, given in (2), is estimated by Monte Carlo. This prior probability depends only on the prior distribution and the number of centers. Recall that the prior distribution in the example was that μ_1 and μ_2 are independent and uniform on $[-4.595,\ 4.595]$, and σ^2 is independently distributed as an inverse gamma distribution. Multiple sets of values of μ_1, μ_2 and σ^2 are sampled from the prior distribution giving a sample of $\mu = \mu_1 - \mu_2$ and and $\omega = \sigma\sqrt{2}$. For each value of μ and ω, the integrand in (2) is calculated for $N = 7$ and then all are averaged. The average of these values, 0.3744, is an estimate of the prior probability of qualitative interaction for this case. Note that, the larger N is, the larger the prior probability of qualitative interaction.

Factors considered in the simulations are the degree of balance of the subjects in subgroups, the magnitude of qualitative interaction and the sample size.

For each specified value of ϕ, corresponding success probabilities p_{ij} are derived: it is arbitrarily assumed that the sum of the success probabilities in each group is 1. For example a log odds ratio of 2.0, is given by $p_1 = 0.269$ and $p_2 = 0.731$. For each combination of ϕ and $\mathbf{n}$, the type I error probabilities and power are estimated empirically for G&S and P&G tests as well as two Bayes test that compares BF_{01} to m for each of $m = \frac{1}{32}$ and $m = \frac{1}{150}$. These values of m correspond to "very strong evidence" with Jeffreys' scale and with Kass and Raftery's scale respectively. All of the four tests, the two based on BF_{01} as well as the G&S and P&G tests, are applied to each data set. The empirical type I error probability and power for different tests are compared to each other. In practice, the exact value of the BF_{01} should be reported to quantify the evidence for qualitative interaction, but to make a comparison with the G&S and P&G tests a cut-off value is used. The impact of different combinations of ϕ and $\mathbf{n}$ is examined in the simulations. The impact of balance was a concern among the investigators in our motivating example and is also examined.

Because the data are simulated from a model with the p_{ij} fixed these are power and size calculations in the frequentist framework: and the power of the BF_{01} test is the frequentist

power of the Bayesian procedure.

For each of 5000 simulated data sets at each combination of ϕ and $\mathbf{n}$, the LRT is used to test for any interaction by fitting two fixed effects generalized linear models (GLM's), one with main effects only and the second with interaction and comparing the difference in deviance to a χ^2_6 distribution. The G&S and P&G are both recommended to be used protected by the LRT: that is they are only declared significant if both the qualitative interaction test and the LRT is significant at 0.05. Both the protected and the unprotected tests were examined in the simulations. The posterior probability of qualitative interaction was calculated irrespectively of the result of the LRT. The data sets are generated in R [25], and G&S and P&G tests are applied in R. The data set is passed passed to WinBUGS [26] using R2WinBUGS [27] for the posterior calculations and the results returned to R where equation (3) is used to calculate BF_{01}.

3.2. Type I error Rate

To examine the type I error probability, data are generated under several settings corresponding to the absence of qualitative interaction. In the motivating example, it was anticipated that there would be no between center variability: ϕ_j is constant for $j = 1, \ldots, 7$. Seven different sets of values for the ϕ_j for $j = 1, \ldots, N$, and corresponding p_{ij}, were chosen which satisfy the null hypothesis of no qualitative interaction. The total sample size was assumed to be either 65, 130, or 260. Results are given in the first part of Table V in section I.1 of Appendix and also in Figure 1 for a total sample size of 260, and the approximately balanced allocation. In these cases, as in all other cases, both the protected G&S and P&G tests were run on the same simulated data sets. The protected procedure gave almost identical results to the unprotected tests and so only the protected results are shown. The BF_{01} tests were not protected by the LRT. Figure 1 shows that in all cases except the first, the Type I error rates are almost zero. The first case corresponds to $\underline{\phi} = (3.0, 3.0, 0.5, 0.5, 0.5, 0.5, 0.5)$ and BF_{01} was less than $\frac{1}{32}$ in just under 2% of cases. Of the seven values chosen to evaluate the null hypothesis, the first case is closest to the boundary of the region where the null hypothesis is satisfied.

Additional simulations where some of the ϕ_j were set to zero were done, and in some cases the Type I error rate of the Bayesian test was very large, close to 50%. Results of these simulations can be found in Tables VI, VII and VIII in section I.1 of the Appendix. Because of the large Type I error rate in some cases, additional simulations were run protecting the Bayesian test with the test for a main effect of the treatment, using a fixed effects GLM. These results can be found in Table IX in section I.2 of the Appendix. In some, but not all cases, this protection led to more conservative Type I error rates, below the nominal 5% of the frequentist tests. For example if $\underline{\phi} = (\phi, \phi, 0, 0, 0, 0, 0)$, then the unprotected Type I error rate for the $BF_{01} < \frac{1}{32}$ is 0.61 and 0.51 for $\phi = 1$ and $\phi = 2.5$ respectively: when ϕ is smaller the null hypothesis is more likely to be rejected. Using the corresponding protected test the Type I error rate is reduced to 0.08 for $\phi = 1$ but 0.49 for $\phi = 2.5$: when ϕ is larger the test for main effects is almost always rejected and so there is not much difference between protected and unprotected procedures in the Type I error rate.

As might be expected the Type I error rate is large near the boundary of the region in the parameter space of no qualitative interaction. This is the primary difference between the tests based on the BF_{01} and the tests of P&G and G&S: the Type I error probability is large on or near the boundary. Protecting the test based on BF_{01} does not satisfactorily address the large

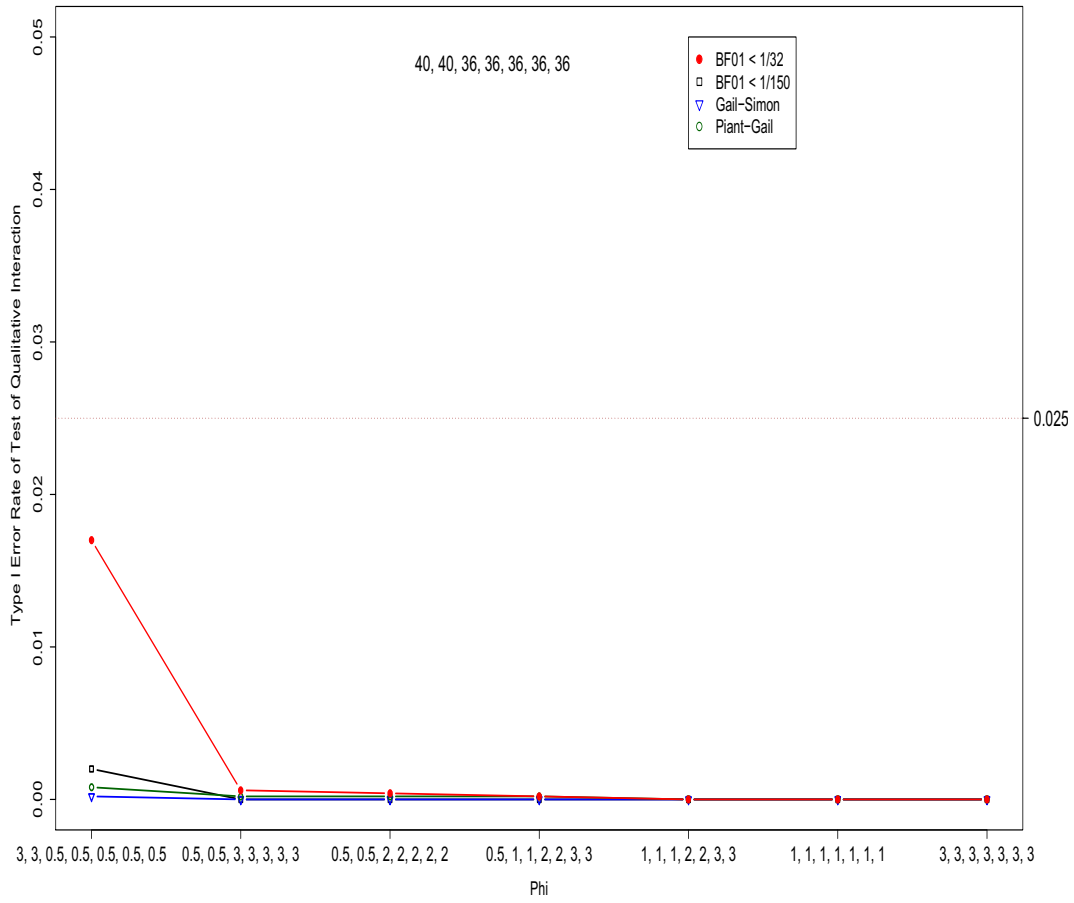

Figure 1. Type I error rates under different scenarios for the most balanced design

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679$, $p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818$, $p_2 = 0.182$, for negative ϕ's p_1 and p_2 are reversed.

Type I error rate in these cases. Section 3.5 shows how this can be addressed by constructing a test with exact expected Type I error probability, where the expectation is over the prior distribution restricted to the space of no qualitative interaction. A similar approach can be used for any test and is done in Section 3.5 for the P&G and G&S tests.

3.3. Power when Treatment is Harmful in One Center

To examine the empirical power, first, the treatment effects are all assumed to be identical in magnitude with exactly one having a negative sign. This reflects a concern that if any one

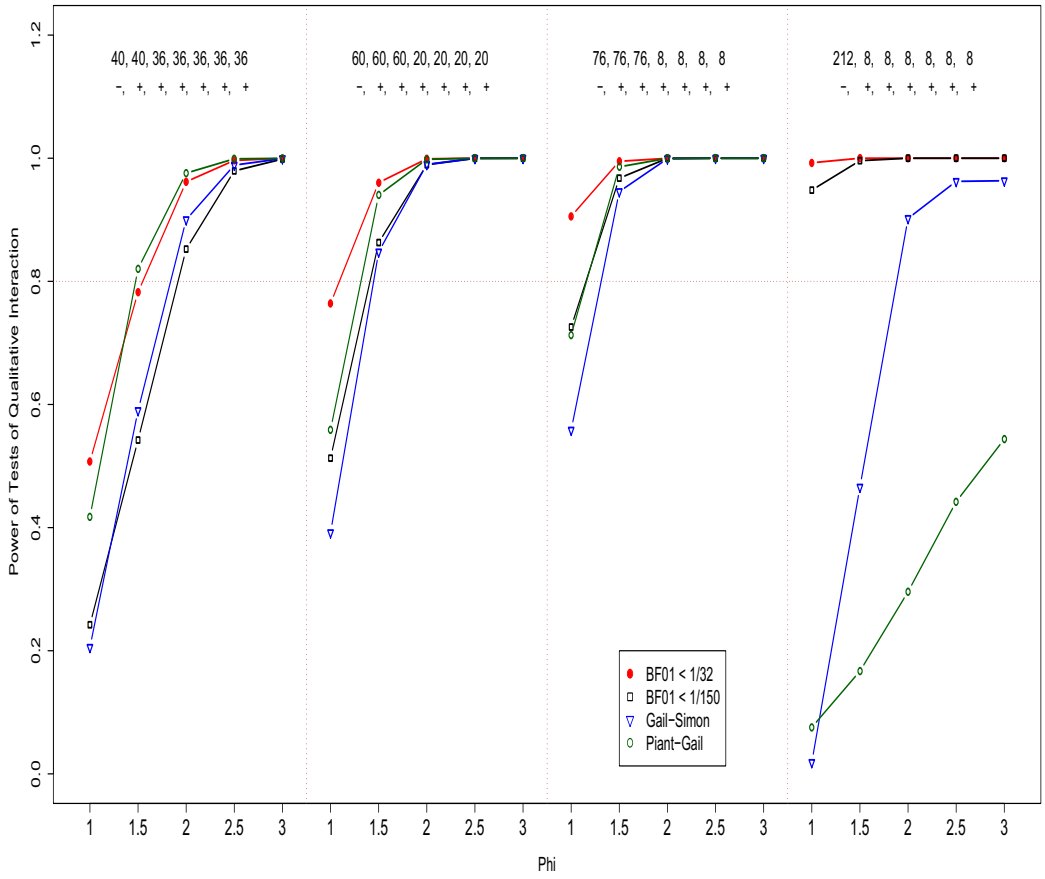

Figure 2. Power of qualitative interaction tests, treatment effect is harmful in the first center, Bayesian unprotected, frequentist protected by LRT

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679$, $p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818$, $p_2 = 0.182$, for negative ϕ's p_1 and p_2 are reversed.

center's laboratory was not effective in isolating the cells for transplantation, the poor results for that center on the experimental arm might adversely impact the trial. Table X provides the results for a total sample size of 260 and 4 designs ranging from most balanced to extremely unbalanced. These results are plotted in Figure 2. This figure shows that when the G&S and P&G tests have power over 80% (in the more balanced cases) the BF_{01} tests have slightly higher power. When the G&S and P&G tests have low power (in the most unbalanced cases) the BF_{01} tests have high power.

In Figure 3 the effect of change from balanced to unbalanced designs is inspected. This figure shows the power for all four combinations of $n.130/520$ and balance / imbalance. The tests

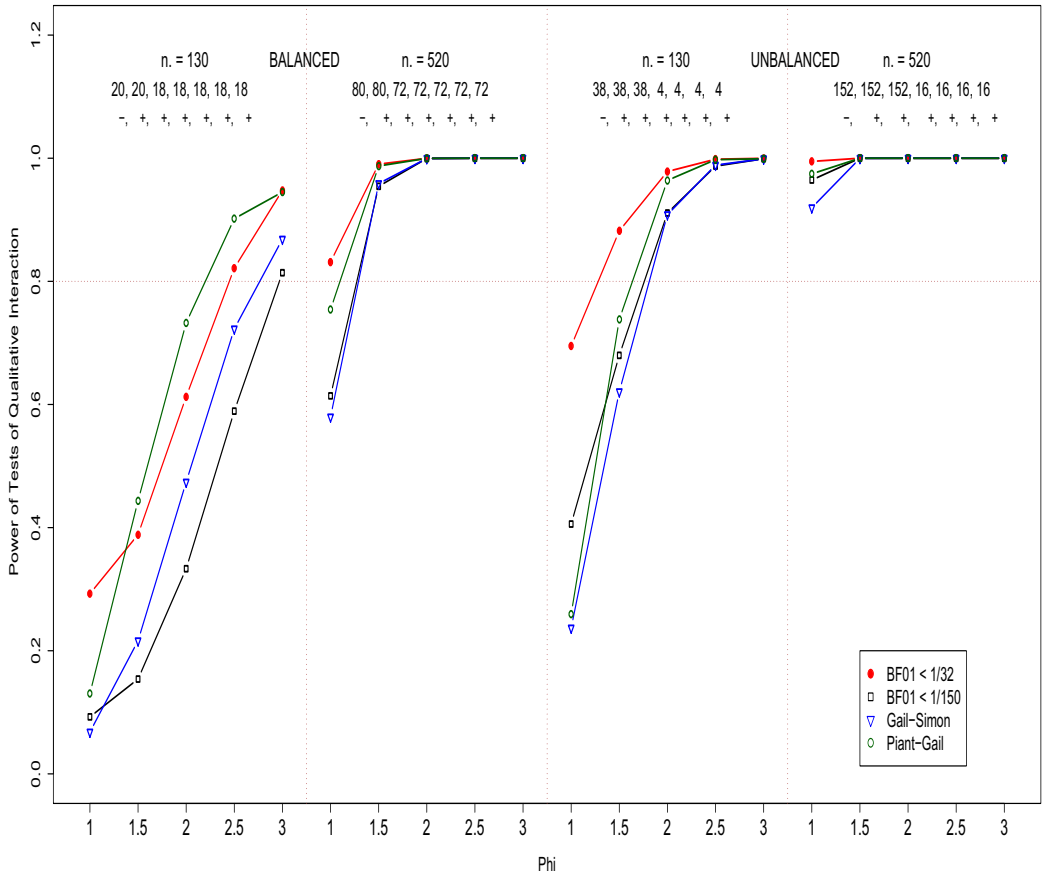

Figure 3. The effect of total sample size on the power of qualitative interaction tests for the balanced and unbalanced designs, treatment effect is harmful in the first center, Bayesian unprotected, frequentist protected

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679$, $p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818$, $p_2 = 0.182$, for negative ϕ's p_1 and p_2 are reversed.

based on BF_{01} have more power than other tests when the design is unbalanced. This finding is consistently observed in other cases. Simulation results used in this figure can be found in Table XI.

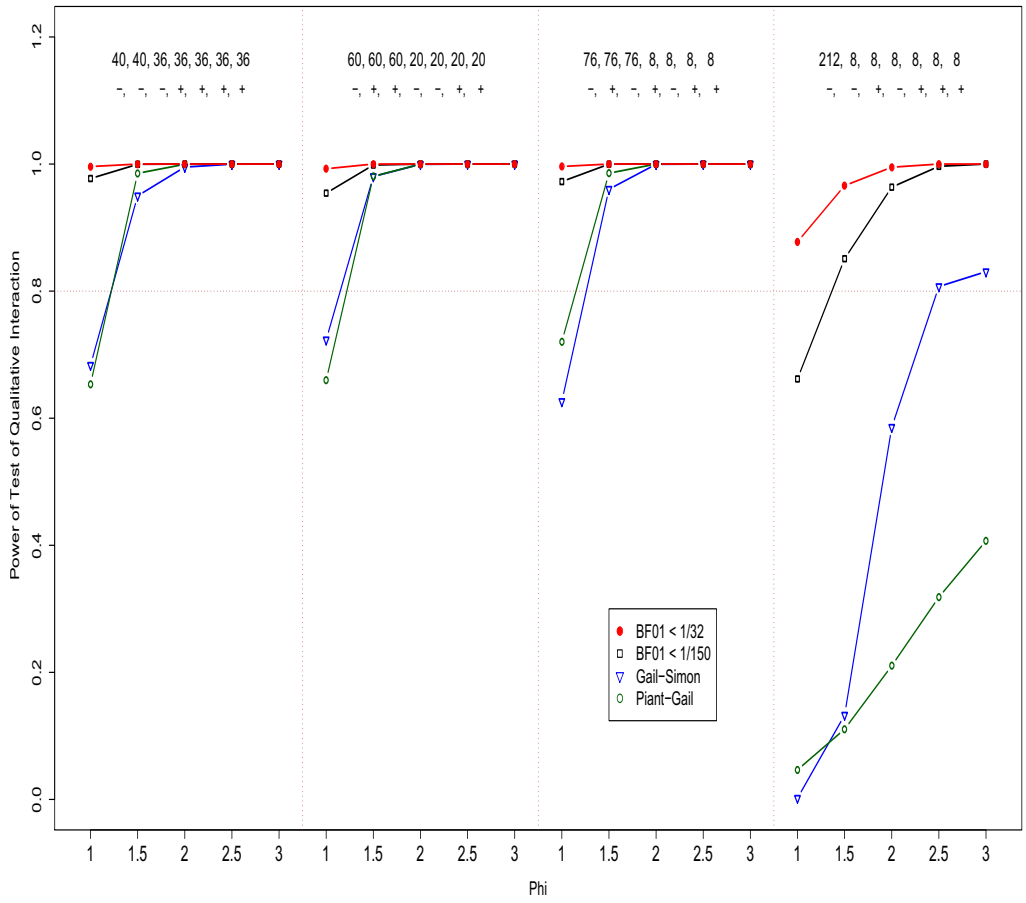

Figure 4. Power of qualitative interaction tests, treatment effect is harmful in three centers

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679$, $p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818$, $p_2 = 0.182$, for negative ϕ's p_1 and p_2 are reversed.

3.4. Power when Treatment is Harmful in Three Centers

Table XII and Figure 4 provide corresponding results when the treatment is beneficial in four centers and harmful in three. The pattern is similar, but with higher power overall. Simulations show that tests based on BF_{01} have higher power than tests of G&S and P&G in this case.

The P&G test is powerful for balanced designs, but as balance declines, the power of this test decreases dramatically. P&G [11] claimed that the P&G test is more powerful when the treatment effect is harmful in only a few subsets and beneficial in most of the subsets. This claim is confirmed for the balanced designs, but not for the unbalanced ones.

3.5. Bayes Test with Exact Size, Three Centers

In this section an example is given of how, for a specified design and prior distribution, a test of significance with exact Type I error probability can be constructed. The example assumes 40 subjects in each treatment in each of three centers. The prior distribution is assumed to be that from the example of Section 2.1: the between center variance has an inverse gamma distribution with parameters $\alpha = 2$ and $\beta = 1.5$ and both μ_1 and μ_2 have a uniform distribution on the interval [-4.595, 4.595], and all three hyperparameters are independent. Define π as the value of the prior probability defined in Equation (2): for this design and prior distribution $\pi = 0.24$. From the prior distribution, 10^5 sets of values of p_{ij}, for $i = 1, 2$ and $j = 1, 2, 3$, are generated to give 10^5 sets $\underline{\phi} = (\phi_1, \phi_2, \phi_3)^T$. QI occurs when there is at least one negative ϕ_j and at least one positive $\phi_{j'}$, where $j \neq j'$: a proportion π of the 10^5 values ϕ are expected to correspond to this alternative hypothesis.

For each of the 10^5 sets of p_{ij}, data are generated and the posterior BF_{01} is calculated. Similarly, G&S and P&G statistics are calculated for each data set. For the data sets corresponding to $\underline{\phi}$ in the null hypothesis region, each of the three sets of simulated test statistics are ranked and the critical values that give exact tests of size 0.05 are estimated. The simulated test statistics that correspond to $\underline{\phi}$ lying in the alternative space, are also ranked and the estimated power is the proportion of them that exceed the critical value.

The three critical values for this example are such that, for a test of size 0.05, the null hypothesis of no qualitative interaction is rejected if BF_{01} is less than 0.23, the test statistic of G&S is greater than 1.9×10^{-7}, and the test statistic of P&G is greater than 0.0004. The power of each exact test is 74%, 62% and 73% for the BF_{01}, $G\&S$ and $P\&G$ methods respectively. This simulation procedure has therefore led to exact tests for QI. The simulations were repeated for an unbalanced design (60, 40 and 20 subjects in each subgroup in each trial) and corresponding cut-off values are recalculated. The estimated power is very similar to the power for balanced design for all three tests: 73%, 60%, 70% respectively.

4. CONCLUSIONS and DISCUSSION

This paper has developed an approach to quantifying the evidence for QI, where QI can be defined from the context (for example as clinically meaningful QI, as in the IHAST example). In the frequentist framework, significance testing forces a decision to reject or not reject the null hypothesis, but quantifying the evidence may be more important as subgroup findings are typically viewed with scepticism and are typically required to be verified [6]. In addition, a frequentist test procedure can be based on the Bayes Factor, and often has better power in the examples examined than the existing tests of G&S and P&G, particularly in unbalanced cases, or when some subgroups have a small sample size. In general, in the examples simulated, if qualitative interaction is big enough to detect with the frequentist test, it does not matter much which test to use.

The Bayesian approach developed here requires the specification of a prior distribution. The prior distribution used in this paper assumes that the treatment responses in each subgroup are exhangeable: this reflects the belief that response to treatment in each subgroup is potentially different, but probably similar and therefore QI is unlikely. Alternative prior distributions could be used, in particular one where treatment effects were exchangeable between subgroups

might be appropriate in some situations. The G&S and P&G statistics are based on a fixed effect model and do not incorporate a prior belief that QI is unlikely.

Note also, a claim of P&G that their test should be more powerful than the test of G&S when a treatment effect is of opposite sign from most in only a few of the subgroup categories, was not verified in the simulations examined here. The test based on G&S typically has more power than or similar power to the power of the test of P&G in the examples examined here with binary responses.

For balanced designs in the cases examined, the G&S and P&G tests perform very similarly, but for unbalanced designs, the G&S test has higher power. The test based on BF_{01} appears to be particularly powerful in unbalanced cases, although in some cases this is at the expense of increased Type I error probability. In the case when the log odds ratio in each subgroup, ϕ_j, are identical and there is no treatment by subgroup interaction, all four tests are highly conservative. These two issues can be adressed through constructing an exact test based on the BF_{01}, and an example is provided in Section 3.5, but for each combination of prior distribution and design simulations are required and this procedure has not been explored further.

A limitation of the examples studied in this paper is that they are specific to the choice of prior distribution and design. For any one clinical trial, however, properties of the procedures, and the different tests, can be examined before the results of the trial are available.

5. ACKNOWLEDGEMENT

We are thankful to Professor George Woodworth for giving helpful comments and suggesting the exact tests, and to Professor Michael Todd for providing the IHAST data.

REFERENCES

1. Dixon DO, Simon R. Bayesian Subset Analysis in a Colorectal Cancer Clinical Trial. *Statistics in Medicine* 1992; **11**:13–22.
2. Berry DA. Subgroup Analyses. *Biometrics* 1990; **4**:1227–1230.
3. Simon R. Bayesian Subset Analysis: Application to studying treatment-by-gender interactions. *Statistics in Medicine* 2002; **21**:2909–2916.
4. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis)uses of baseline data in clinical trials. *The Lancet* 2000; **335**:1064 – 1069.
5. Pocock SJ, Assmann SF, Enos LE, Kasten LE. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine* 2002; **21**:2917–2930.
6. Follmann D. Subgroups and Interactions. In *Advances in Clinical Trials Biostatistics,* Geller NL (editor). CRC, 2003; 124 – 142.
7. Peto R. Statistical Aspects of Cancer Trials. In *Treatment of Cancer,* Halnan KE (editor). Chapmann and Hall, 1982; 867–871.
8. Gail M. and Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985; **41**:361–372.
9. Silvapulle MJ. Tests Against Qualitative Interaction: Exact critical Values and Robust Tests. *Biometrics* 2001; **57**:1157–1165.
10. Russek-Cohen E, Simon RM. Qualitative Interactions in Multifactor Studies. *Biometrics* 1993; **49**:467–477.
11. Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine* 1993; **12**:1239 – 1248.
12. Li J, Chan ISF. Detecting Qualitative Interactions in Clinical Trials: An Extension of Range Test. *Journal of Biopharmaceutical Statistics* 2006; **16**:831–841.

13. Pan G, Wolfe D. Test for qualitative interaction of clinical significance. *Statistics in Medicine* 1997; **16**:1645–1652.
14. Yan X, Sun X. Testing for qualitative interaction. *Encyclopedia of Biopharmaceutical Statistics* 2005; **1**:1–8.
15. Ciminera JL, Heyse JF, Nguyen HH, Tukey JW. Tests for Qualitative Treatment-by-Centre Interaction Using a 'Pushback' Procedure. *Statistics in Medicine* 1993; **12**:1033–1045.
16. Boos DD, Brownie C, Zhang J. Estimating the Magnitude of Interaction. *Institute of Statistics Mimeo Series, 2285* 1996; North Carolina State University, Raleigh, NC.
17. Wiens BL, Heyse JF. Testing for Interaction in Studies of Noninferiority. *Journal of Biopharmaceutical Statistics* 2003; **13**(1):103–115.
18. Duncan DB. A Bayesian Approach to Multiple Comparisons. *Technometrics* 1965; **7**(2):171–222.
19. Dixon DO, Simon R. Bayesian Subset Analysis. *Biometrics* 1991; **47**:871–881.
20. White IR, Pocock SJ, Wang D. Eliciting and using expert opinions about influence of patient characteristics on treatment effects: a Bayesian analysis of the CHARM trials. *Statistics in Medicine* 2005; **24**:3805–3821.
21. Good IJ. *Good Thinking: The Foundations of Probability and Its Applications* University of Minnesota Press: Minneapolis, 1983;
22. Kass RE, Raftery AE. Bayes Factors. *Journal of the American Statistical Association* 1995; **90** (430):773–795.
23. Jeffreys H. *Theory of Probability* (3rt edn), Oxford University Press: Oxford, 1961;
24. Todd MM, Hindman BJ, Clarke WR, Torner JC. Mild Intraoperative Hypothermia during Surgery for Intracranial Aneurysm. *New England Journal of Medicine* 2005; **352** (2):135–145.
25. R Development Core Team R: A language and environment for statistical computing, Vienna, Austria *R Foundation for Statistical Computing* 2005; **ISBN: 3-900051-07-0, http://www.R-project.org**
26. Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual, 1.4.1 *Medical Research Council (MRC)* 2003; **ISBN: 3-900051-07-0, www.mrc-bsu.cam.ac.uk/bugs**
27. Sturtz S, Ligges U, Gelman A. R2WinBUGS: A Package for Running WinBUGS from R. *Journal of Statistical Software, http://www.jstatsoft.org/v12/i03* 2005; **12** (3):1–16.
28. Bayman EO. Bayesian hierarchical models for multi-center clinical trials: power and subgoup analysis. *Ph.D. dissertation* 2008 The University of Iowa, Iowa City, IA; **1** – 218.

APPENDIX

Simulation Results for Section 3

I.1. Type I Error Rates

Table V. Type I error rates for test of qualitative interaction for different designs

Test	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7
			n_{ij}**: 40, 40, 36, 36, 36, 36, 36**				
$BF_{01} < 1/32$	0.02	0.00	0.00	0.00	0.00	0.00	0.00
$BF_{01} < 1/150$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G&S Unprot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Any Int LRT	1.00	1.00	0.83	0.97	0.91	0.06	0.07
G&S Prot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			n_{ij}**: 60, 60, 60, 20, 20, 20, 20**				
$BF_{01} < 1/32$	0.02	0.00	0.00	0.00	0.00	0.00	0.00
$BF_{01} < 1/150$	0.01	0.00	0.00	0.00	0.00	0.00	0.00
G&S Unprot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Any Int LRT	1.00	1.00	0.89	0.85	0.93	0.06	0.08
G&S Prot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			n_{ij}**: 76, 76, 76, 8, 8, 8, 8**				
$BF_{01} < 1/32$	0.02	0.00	0.00	0.00	0.00	0.00	0.00
$BF_{01} < 1/150$	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G&S Unprot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Any Int LRT	1.00	1.00	0.91	0.60	0.74	0.09	0.11
G&S Prot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
			n_{ij}**: 212, 8, 8, 8, 8, 8, 8**				
$BF_{01} < 1/32$	0.04	0.00	0.00	0.00	0.00	0.01	0.00
$BF_{01} < 1/150$	0.01	0.00	0.00	0.00	0.00	0.00	0.00
G&S Unprot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Any Int LRT	0.95	0.98	0.67	0.62	0.83	0.10	0.13
G&S Prot	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679, p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818, p_2 = 0.182$. $\phi_1 = 3, 3, 0.5, ..., 0.5$, $\phi_2 = 0.5, 0.5, 3, ..., 3$, $\phi_3 = 0.5, 0.5, 2, ..., 2$, $\phi_4 = 0.5, 1, 1, 2, 2, 3, 3$, $\phi_5 = 1, 1, 1, 2, 2, 3, 3$, $\phi_6 = 1, ..., 1$, $\phi_7 = 3, ..., 3$.

Table VI. Type I error rates for tests of qualitative interaction for balanced designs

Test	$\phi = 1$	$\phi = 1.5$	$\phi = 2$	$\phi = 2.5$	$\phi = 3$
		n_{ij}**: 40,40,36,36,36,36,36**			
		ϕ, ϕ**, 0, 0, 0, 0, 0**			
$BF_{01} < 1/32$	0.61	0.54	0.51	0.51	0.52
$BF_{01} < 1/150$	0.27	0.24	0.22	0.22	0.24
G&S Unprot	0.02	0.03	0.03	0.03	0.03
P&G Unprot	0.02	0.03	0.03	0.03	0.03
Any Int LRT	0.47	0.84	0.99	1.00	1.00
G&S Prot	0.02	0.03	0.03	0.03	0.03
P&G Prot	0.02	0.03	0.03	0.03	0.03
		n_{ij}**: 40,40,36,36,36,36,36**			
		0, 0, $\phi, \phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	0.09	0.05	0.04	0.04	0.04
$BF_{01} < 1/150$	0.02	0.01	0.01	0.01	0.01
G&S Unprot	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.02	0.02	0.02	0.02	0.02
Any Int LRT	0.48	0.86	0.99	1.00	1.00
G&S Prot	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.02	0.02	0.00	0.02	0.02
		n_{ij}**: 60,60,60,20,20,20,20**			
		ϕ, ϕ**, 0, 0, 0, 0, 0**			
$BF_{01} < 1/32$	0.52	0.46	0.44	0.44	0.45
$BF_{01} < 1/150$	0.22	0.19	0.17	0.18	0.19
G&S Unprot	0.02	0.03	0.03	0.03	0.03
P&G Unprot	0.03	0.04	0.04	0.04	0.04
Any Int LRT	0.53	0.90	0.99	1.00	1.00
G&S Prot	0.02	0.03	0.03	0.03	0.03
P&G Prot	0.03	0.04	0.04	0.04	0.04
		n_{ij}**: 60,60,60,20,20,20,20**			
		0, 0, $\phi, \phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	0.14	0.08	0.06	0.06	0.06
$BF_{01} < 1/150$	0.03	0.01	0.01	0.01	0.01
G&S Unprot	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.01	0.01	0.01	0.00	0.01
Any Int LRT	0.54	0.91	1.00	1.00	1.00
G&S Prot	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.01	0.01	0.01	0.00	0.01

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679, p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818, p_2 = 0.182$.

Table VII. Type I error rates for tests of qualitative interaction for unbalanced designs

Test	$\phi = 1$	$\phi = 1.5$	$\phi = 2$	$\phi = 2.5$	$\phi = 3$
		n_{ij}**: 76,76,76,8,8,8,8**			
		ϕ, ϕ**, 0, 0, 0, 0, 0**			
$BF_{01} < 1/32$	0.39	0.32	0.29	0.29	0.30
$BF_{01} < 1/150$	0.13	0.10	0.09	0.09	0.10
G&S Unprot	0.01	0.01	0.01	0.01	0.01
P&G Unprot	0.01	0.01	0.01	0.01	0.01
Any Int LRT	0.54	0.90	0.99	1.00	1.00
G&S Prot	0.01	0.01	0.01	0.01	0.01
P&G Prot	0.01	0.01	0.01	0.01	0.01
		n_{ij}**: 76,76,76,8,8,8,8**			
		0, 0, $\phi, \phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	0.25	0.14	0.11	0.09	0.09
$BF_{01} < 1/150$	0.07	0.04	0.03	0.02	0.02
G&S Unprot	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.01	0.02	0.02	0.02	0.02
Any Int LRT	0.57	0.91	0.99	1.00	1.00
G&S Prot	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.00	0.02	0.02	0.02	0.02
		n_{ij}**: 212,8,8,8,8,8,8**			
		ϕ, ϕ**, 0, 0, 0, 0, 0**			
$BF_{01} < 1/32$	0.39	0.31	0.27	0.24	0.25
$BF_{01} < 1/150$	0.15	0.10	0.08	0.07	0.08
G&S Unprot	0.01	0.01	0.01	0.01	0.01
P&G Unprot	0.00	0.01	0.00	0.01	0.01
Any Int LRT	0.32	0.62	0.85	0.97	1.00
G&S Prot	0.01	0.01	0.01	0.01	0.01
P&G Prot	0.00	0.01	0.00	0.01	0.01
		n_{ij}**: 212,8,8,8,8,8,8**			
		0, 0, $\phi, \phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	0.26	0.13	0.08	0.06	0.05
$BF_{01} < 1/150$	0.04	0.03	0.02	0.01	0.01
G&S Unprot	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.00	0.00	0.00	0.01	0.00
Any Int LRT	0.37	0.68	0.90	0.98	1.00
G&S Prot	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.00	0.00	0.00	0.01	0.00

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679, p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818, p_2 = 0.182$.

Table VIII. Type I error rates for tests of qualitative interaction

Test	$\phi = 1$	$\phi = 1.5$	$\phi = 2$	$\phi = 2.5$	$\phi = 3$
		n_{ij}**: 40,40,36,36,36,36,36**			
		ϕ, ϕ, ϕ, ϕ**, 0, 0, 0**			
$BF_{01} < 1/32$	0.20	0.13	0.12	0.12	0.13
$BF_{01} < 1/150$	0.05	0.03	0.03	0.03	0.03
G&S Unprot	0.01	0.01	0.01	0.01	0.01
P&G Unprot	0.01	0.02	0.02	0.02	0.02
Any Int LRT	0.53	0.90	0.99	1.00	1.00
G&S Prot	0.01	0.01	0.01	0.01	0.01
P&G Prot	0.01	0.02	0.02	0.02	0.02
		n_{ij}**: 60,60,60,20,20,20,20**			
		ϕ, ϕ, ϕ, ϕ**, 0, 0, 0**			
$BF_{01} < 1/32$	0.13	0.08	0.07	0.07	0.07
$BF_{01} < 1/150$	0.03	0.01	0.01	0.01	0.01
G&S Unprot	0.01	0.01	0.01	0.01	0.01
P&G Unprot	0.02	0.02	0.02	0.02	0.02
Any Int LRT	0.40	0.76	0.96	1.00	1.00
G&S Prot	0.01	0.01	0.01	0.01	0.01
P&G Prot	0.02	0.02	0.02	0.02	0.02
		n_{ij}**: 76,76,76,8,8,8,8**			
		ϕ, ϕ, ϕ, ϕ**, 0, 0, 0**			
$BF_{01} < 1/32$	0.06	0.02	0.02	0.01	0.02
$BF_{01} < 1/150$	0.01	0.00	0.00	0.00	0.00
G&S Unprot	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.00	0.00	0.00	0.00	0.00
Any Int LRT	0.21	0.42	0.66	0.85	0.95
G&S Prot	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.00	0.00	0.00	0.00	0.00
		n_{ij}**: 212,8,8,8,8,8,8**			
		ϕ, ϕ, ϕ, ϕ**, 0, 0, 0**			
$BF_{01} < 1/32$	0.17	0.08	0.05	0.04	0.04
$BF_{01} < 1/150$	0.04	0.02	0.01	0.01	0.01
G&S Unprot	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.00	0.00	0.00	0.00	0.00
Any Int LRT	0.24	0.45	0.68	0.87	0.96
G&S Prot	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.00	0.00	0.00	0.00	0.00

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679, p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818, p_2 = 0.182$.

I.2. Type I Error Rates when Bayesian Test is Protected by the Overall Test

Table IX. Type I error rates when Bayesian test is protected with the overall test: GLM trt effect p-val < 0.05

Test	ϕ **= 1**	ϕ **= 1.5**	ϕ **= 2**	ϕ **= 2.5**	ϕ **= 3**
		n_{ij}**: 40,40,36,36,36,36,36**			
		ϕ, ϕ**, 0, 0, 0, 0, 0**			
$BF_{01} < 1/32$	0.08	0.26	0.43	0.49	0.53
$BF_{01} < 1/150$	0.01	0.05	0.15	0.21	0.24
G&S Unprot	0.02	0.03	0.03	0.03	0.03
P&G Unprot	0.02	0.03	0.03	0.03	0.04
Any Int LRT	0.46	0.84	0.98	1.00	1.00
G&S Prot	0.02	0.03	0.03	0.03	0.03
P&G Prot	0.02	0.03	0.03	0.03	0.04
		n_{ij}**: 40,40,36,36,36,36,36**			
		0, 0, $\phi, \phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	0.07	0.05	0.04	0.04	0.04
$BF_{01} < 1/150$	0.01	0.01	0.01	0.01	0.01
G&S Unprot	0.00	0.00	0.00	0.00	0.00
P&G Unprot	0.02	0.02	0.02	0.02	0.02
Any Int LRT	0.47	0.86	0.99	1.00	1.00
G&S Prot	0.00	0.00	0.00	0.00	0.00
P&G Prot	0.02	0.02	0.02	0.02	0.02

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679, p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818, p_2 = 0.182$. $\phi_1 = 3, 3, 0.5, ..., 0.5$, $\phi_2 = 0.5, 0.5, 3, ..., 3$, $\phi_3 = 0.5, 0.5, 2, ..., 2$, $\phi_4 = 0.5, 1, 1, 2, 2, 3, 3$, $\phi_5 = 1, 1, 1, 2, 2, 3, 3$, $\phi_6 = 1, ..., 1$, $\phi_7 = 3, ..., 3$.

I.3. Difference in One Center

Table X. Power results for test of qualitative interaction where treatment effect is harmful in the first center and beneficial in the rest of the centers

Test	$\phi = 1$	$\phi = 1.5$	$\phi = 2$	$\phi = 2.5$	$\phi = 3$
		n_{ij}**: 40,40,36,36,36,36,36**			
		$-\phi, \phi, \phi, \phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	0.51	0.78	0.96	1.00	1.00
$BF_{01} < 1/150$	0.24	0.54	0.85	0.98	1.00
G&S Unprot	0.21	0.59	0.90	0.99	1.00
P&G Unprot	0.42	0.82	0.98	1.00	1.00
Any Int LRT	0.90	1.00	1.00	1.00	1.00
G&S Prot	0.21	0.59	0.90	0.99	1.00
P&G Prot	0.42	0.82	0.98	1.00	1.00
		n_{ij}**: 60,60,60,20,20,20,20**			
		$-\phi, \phi, \phi, \phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	0.76	0.96	1.00	1.00	1.00
$BF_{01} < 1/150$	0.51	0.86	0.99	1.00	1.00
G&S Unprot	0.39	0.84	0.99	1.00	1.00
P&G Unprot	0.56	0.94	1.00	1.00	1.00
Any Int LRT	0.97	1.00	1.00	1.00	1.00
G&S Prot	0.39	0.84	0.99	1.00	1.00
P&G Prot	0.56	0.94	1.00	1.00	1.00
		n_{ij}**: 76,76,76,8,8,8,8**			
		$-\phi, \phi, \phi, \phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	0.91	1.00	1.00	1.00	1.00
$BF_{01} < 1/150$	0.73	0.97	1.00	1.00	1.00
G&S Unprot	0.56	0.95	1.00	1.00	1.00
P&G Unprot	0.71	0.99	1.00	1.00	1.00
Any Int LRT	0.99	1.00	1.00	1.00	1.00
G&S Prot	0.56	0.95	1.00	1.00	1.00
P&G Prot	0.71	0.99	1.00	1.00	1.00
		n_{ij}**: 212,8,8,8,8,8,8**			
		$-\phi, \phi, \phi, \phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	0.93	1.00	1.00	1.00	1.00
$BF_{01} < 1/150$	0.91	1.00	1.00	1.00	1.00
G&S Unprot	0.01	0.46	0.90	0.97	0.97
P&G Unprot	0.07	0.17	0.32	0.46	0.57
Any Int LRT	0.94	1.00	1.00	1.00	1.00
G&S Prot	0.01	0.46	0.90	0.97	0.97
P&G Prot	0.07	0.17	0.32	0.46	0.57

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679, p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818, p_2 = 0.182$. $\phi_1 = 3, 3, 0.5, ..., 0.5$, $\phi_2 = 0.5, 0.5, 3, ..., 3$, $\phi_3 = 0.5, 0.5, 2, ..., 2$, $\phi_4 = 0.5, 1, 1, 2, 2, 3, 3$, $\phi_5 = 1, 1, 1, 2, 2, 3, 3$, $\phi_6 = 1, ..., 1$, $\phi_7 = 3, ..., 3$.

Table XI. Power results for test of qualitative interaction where treatment effect is harmful in the first center and beneficial in the rest of the centers, the effect of sample size on power

Design	**Test**	$\phi = 1$	$\phi = 1.5$	$\phi = 2$	$\phi = 2.5$	$\phi = 3$
		n_{ij}**: 20,20,18,18,18,18,18**				
		$-\phi, \phi, \phi, \phi, \phi, \phi, \phi$				
$BF_{01} < 1/32$	0.29	0.39	0.61	0.82	0.95	
$BF_{01} < 1/150$	0.09	0.15	0.33	0.59	0.81	
G&S Unprot	0.07	0.22	0.47	0.72	0.87	
P&G Unprot	0.13	0.44	0.73	0.90	0.95	
Any Int LRT	0.59	0.93	1.00	1.00	1.00	
G&S Prot	0.07	0.22	0.47	0.72	0.87	
P&G Prot	0.13	0.44	0.73	0.90	0.95	
		n_{ij}**: 80, 80, 72, 72, 72, 72, 72**				
		$-\phi, \phi, \phi, \phi, \phi, \phi, \phi$				
$BF_{01} < 1/32$	0.83	0.99	1.00	1.00	1.00	
$BF_{01} < 1/150$	0.61	0.95	1.00	1.00	1.00	
G&S Unprot	0.58	0.96	1.00	1.00	1.00	
P&G Unprot	0.75	0.99	1.00	1.00	1.00	
Any Int LRT	1.00	1.00	1.00	1.00	1.00	
G&S Prot	0.58	0.96	1.00	1.00	1.00	
P&G Prot	0.75	0.99	1.00	1.00	1.00	
		n_{ij}**: 38, 38, 38, 4, 4, 4, 4**				
		$-\phi, \phi, \phi, \phi, \phi, \phi, \phi$				
$BF_{01} < 1/32$	0.69	0.88	0.98	1.00	1.00	
$BF_{01} < 1/150$	0.41	0.68	0.91	0.99	1.00	
G&S Unprot	0.24	0.62	0.91	0.99	1.00	
P&G Unprot	0.26	0.74	0.96	1.00	1.00	
Any Int LRT	0.85	0.99	1.00	1.00	1.00	
G&S Prot	0.24	0.62	0.91	0.99	1.00	
P&G Prot	0.26	0.74	0.96	1.00	1.00	
		n_{ij}**: 152, 152, 152, 16, 16, 16, 16**				
		$-\phi, \phi, \phi, \phi, \phi, \phi, \phi$				
$BF_{01} < 1/32$	1.00	1.00	1.00	1.00	1.00	
$BF_{01} < 1/150$	0.96	1.00	1.00	1.00	1.00	
G&S Unprot	0.92	1.00	1.00	1.00	1.00	
P&G Unprot	0.98	1.00	1.00	1.00	1.00	
Any Int LRT	1.00	1.00	1.00	1.00	1.00	
G&S Prot	0.92	1.00	1.00	1.00	1.00	
P&G Prot	0.98	1.00	1.00	1.00	1.00	

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679, p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818, p_2 = 0.182$. $\phi_1 = 3, 3, 0.5, ..., 0.5$, $\phi_2 = 0.5, 0.5, 3, ..., 3$, $\phi_3 = 0.5, 0.5, 2, ..., 2$, $\phi_4 = 0.5, 1, 1, 2, 2, 3, 3$, $\phi_5 = 1, 1, 1, 2, 2, 3, 3$, $\phi_6 = 1, ..., 1$, $\phi_7 = 3, ..., 3$.

I.4. Difference in Three Centers

Table XII. Power results for test of qualitative interaction where treatment effect is harmful in three centers and beneficial in the rest of the centers

Test	$\phi = 1$	$\phi = 1.5$	$\phi = 2$	$\phi = 2.5$	$\phi = 3$
		n_{ij}**: 40,40,36,36,36,36,36**			
		$-\phi, -\phi, -\phi, \phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	1.00	1.00	1.00	1.00	1.00
$BF_{01} < 1/150$	1.00	1.00	1.00	1.00	1.00
G&S Unprot	0.69	0.95	1.00	1.00	1.00
P&G Unprot	0.65	0.99	1.00	1.00	1.00
Any Int LRT	1.00	1.00	1.00	1.00	1.00
G&S Prot	0.68	0.95	1.00	1.00	1.00
P&G Prot	0.65	0.99	1.00	1.00	1.00
		n_{ij}**: 60,60,60,20,20,20,20**			
		$-\phi, \phi, \phi, -\phi, -\phi, \phi, \phi$			
$BF_{01} < 1/32$	0.99	1.00	1.00	1.00	1.00
$BF_{01} < 1/150$	0.95	1.00	1.00	1.00	1.00
G&S Unprot	0.73	0.99	1.00	1.00	1.00
P&G Unprot	0.67	0.98	1.00	1.00	1.00
Any Int LRT	0.99	1.00	1.00	1.00	1.00
G&S Prot	0.73	0.98	1.00	1.00	1.00
P&G Prot	0.67	0.98	1.00	1.00	1.00
		n_{ij}**: 76,76,76,8,8,8,8**			
		$-\phi, \phi, -\phi, \phi, -\phi, \phi, \phi$			
$BF_{01} < 1/32$	1.00	1.00	1.00	1.00	1.00
$BF_{01} < 1/150$	0.97	1.00	1.00	1.00	1.00
G&S Unprot	0.66	0.96	1.00	1.00	1.00
P&G Unprot	0.73	0.99	1.00	1.00	1.00
Any Int LRT	1.00	1.00	1.00	1.00	1.00
G&S Prot	0.66	0.96	1.00	1.00	1.00
P&G Prot	0.73	0.99	1.00	1.00	1.00
		n_{ij}**: 212,8,8,8,8,8,8**			
		$-\phi, -\phi, \phi, -\phi, \phi, \phi, \phi$			
$BF_{01} < 1/32$	0.88	0.97	1.00	1.00	1.00
$BF_{01} < 1/150$	0.66	0.85	0.96	1.00	1.00
G&S Unprot	0.00	0.13	0.59	0.81	0.83
P&G Unprot	0.05	0.12	0.21	0.32	0.41
Any Int LRT	0.83	1.00	1.00	1.00	1.00
G&S Prot	0.00	0.13	0.59	0.81	0.83
P&G Prot	0.05	0.12	0.21	0.32	0.41

Note: In the simulations success probabilities for $\phi = 1 : p_1 = 0.622, p_2 = 0.378$, for $\phi = 1.5 : p_1 = 0.679, p_2 = 0.321$, for $\phi = 2 : p_1 = 0.731, p_2 = 0.269$, for $\phi = 2.5 : p_1 = 0.777, p_2 = 0.223$, for $\phi = 3 : p_1 = 0.818, p_2 = 0.182$. $\phi_1 = 3, 3, 0.5, ..., 0.5$, $\phi_2 = 0.5, 0.5, 3, ..., 3$, $\phi_3 = 0.5, 0.5, 2, ..., 2$, $\phi_4 = 0.5, 1, 1, 2, 2, 3, 3$, $\phi_5 = 1, 1, 1, 2, 2, 3, 3$, $\phi_6 = 1, ..., 1$, $\phi_7 = 3, ..., 3$.