Research Article

Statistics in Medicine

Received 16 March 2009, Accepted 9 October 2009 Published online in Wiley Interscience

(www.interscience.wiley.com) DOI: 10.1002/sim.3787

Detecting qualitative interaction: A Bayesian approach

Emine Özgür Bayman,[a,b]*† Kathryn Chaloner[b,c] and Mary Kathryn Cowles[c,b]

Differences in treatment effects between centers in a multi-center trial may be important. These differences represent treatment by subgroup interaction. Peto defines qualitative interaction (QI) to occur when the simple treatment effect in one subgroup has a different sign than in another subgroup: this interaction is important. Interaction where the treatment effects are of the same sign in all subgroups is called quantitative and is often not important because the treatment recommendation is identical in all cases. A hierarchical model is used here with exchangeable mean responses to each treatment between subgroups. The posterior probability of QI and the corresponding Bayes factor are proposed as a diagnostic and as a test statistic. The model is motivated by two multi-center trials with binary responses. The frequentist power and size of the test using the Bayes factor are examined and compared with two other commonly used tests. The impact of imbalance between the sample sizes in each subgroup on power 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 the Bayes factor is also suggested assuming the hierarchical model. The Bayes factor provides a concise summary of the evidence for or against QI. It is shown by example that it is easily adapted to summarize the evidence for 'clinically meaningful QI,' defined as the simple effects being of opposite signs and larger in absolute value than a minimal clinically meaningful effect. Copyright © 2009 John Wiley & Sons, Ltd.

Keywords: Bayes factor; Bayesian subgroup analysis; 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 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 subjects into mutually exclusive subsets or subgroups based on values of one or more variables [2]. *Subgroup analysis* is analyzing the treatment effect in each subgroup category.

When two treatments are compared separately for several subgroup categories, and all null hypotheses are true, the probability of making at least one Type I error is greater than the Type I error probability for each category [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 analyses 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 [4]. 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 [5] defines *qualitative* interaction (QI) as arising when the signs of the underlying treatment differences vary across subgroups. When the magnitude of the treatment benefit varies, but the sign does not, then there is a *quantitative* interaction. Quantitative interactions are unimportant and common, whereas QI is important and less likely [5]. QI implies that the subgroup populations should be treated differently and any overall result does not apply to all subgroups.

[a]*Department of Anesthesia, The University of Iowa, Iowa City, IA, U.S.A.*
[b]*Department of Biostatistics, The University of Iowa, Iowa City, IA, U.S.A.*
[c]*Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, IA, U.S.A.*
*Correspondence to: Emine Özgür Bayman, Department of Anesthesia, The University of Iowa, Iowa City, IA, U.S.A.
†*E-mail: emine-bayman@uiowa.edu*

1.1. Review of frequentist methods

The standard test of interaction does not distinguish QI from quantitative interaction. Gail and Simon (G&S) [6] constructed a likelihood ratio test (LRT) to detect QI. 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 accurately estimated, variance. This test is widely used to detect QI. Additional discussion is in [7, 8].

Piantadosi and Gail (P&G) [9] proposed an alternative test of QI 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 the LRT to test for any interaction and then use the test of QI only if the LRT for any interaction is significant. P&G [9] claim 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 [6] 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 [10] extended P&G's range test to use all observed treatment differences rather than only the minimum and the maximum. Pan and Wolfe [11] developed a test of QI with clinical significance which tests whether the difference in treatments is bigger than a pre-defined value. See also [12–14].

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 [4].

In this paper, our Bayesian approach will be compared with the G&S and P&G tests of QI. From a Bayesian perspective, the multiple comparisons issue is different [15, 16], but in examining the properties of a frequentist test based on a Bayes factor the issues are similar.

1.2. Review of Bayesian tests of interaction

In their 1991 paper Dixon and Simon [16] 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. 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 QI is not likely.

In a later paper, Simon [3] proposed a Bayesian approach for subgroup analysis under the assumption of no QI, and used non-informative prior distributions for main effects and informative prior distributions for interaction effects. White *et al.* [17] used a questionnaire to elicit expert beliefs regarding treatment and covariate effects in clinical trials.

The 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 QI is unlikely.

2. The model and definitions

Let the treatment effect in subgroup j be represented by ϕ_j, $j=1,\ldots,N$. Since QI occurs when at least one subgroup's treatment effect is in the opposite direction to that in other subgroups, the appropriate hypotheses are:

$$\begin{aligned} &H_0: \phi_j \geqslant 0 \text{ for all } j, \text{ or, } \phi_j \leqslant 0 \text{ for all } j \\ &H_2: \text{There exists at least one pair } j \neq j' \text{ such that } \phi_j > 0 \text{ and } \phi_{j'} < 0. \end{aligned} \tag{1}$$

Define y_{ijk} as the response of subject k assigned to treatment i in subgroup j in a two group design. Assume that y_{ijk} has an exponential family distribution with mean μ_{ij}. Let θ_{ij} be the result from applying the link function to μ_{ij}, as in a generalized linear model (GLM). Then $\phi_j = \theta_{2j} - \theta_{1j}$ is the simple effect of treatment: if $\phi_j > 0$, the new treatment is superior to the standard therapy in subgroup j. Assume that

$$\phi_j | \mu, \omega^2 \overset{\text{ind}}{\sim} \text{Normal}(\mu, \omega^2), \quad j = 1, \ldots, N. \tag{2}$$

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 QI, Pr(QI), is as follows:

$$\begin{aligned} \Pr(\text{QI}) &= 1 - [\Pr(\phi_1 > 0, \ldots, \phi_N > 0) + \Pr(\phi_1 < 0, \ldots, \phi_N < 0)] \\ &= 1 - \int\int \left[\prod_{j=1}^{N} \Pr(\phi_j | \mu, \omega) > 0 + \prod_{j=1}^{N} \Pr(\phi_j | \mu, \omega) < 0 \right] dp(\mu, \omega) \\ &= \int\int \left\{ 1 - \left[\Phi\left(\frac{\mu}{\omega}\right) \right]^N - \left[1 - \Phi\left(\frac{\mu}{\omega}\right) \right]^N \right\} dp(\mu, \omega). \end{aligned} \tag{3}$$

After specification of a prior distribution for μ and ω, the double integral in Equation (3) may be approximated by the Monte Carlo integration. The posterior probability of QI based on data $\underline{y}$ is denoted by $\Pr(QI|\underline{y})$ and can be computed similarly.

The evidence for or against QI can be quantified by the Bayes factor, BF_{01}. 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 [18, p. 36], who attributes the method to Turing and Jeffreys [19].

Jeffreys [20, p. 432] provided a scale for interpreting the value of the BF_{01} and suggested interpreting BF_{01} in half-units on the $\log_{10}$ scale [19]. 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, as in [21, p. 55]. Kass and Raftery [19] suggested using a more conservative version of Jeffreys' interpretation, where $BF_{01}<\frac{1}{3}, <\frac{1}{20}$ and $<\frac{1}{150}$ are interpreted, respectively, as *positive*, *strong* and *very strong* evidence against the null hypothesis. Note that using BF_{01} to summarize the evidence for QI is different from a test of hypothesis which 'rejects' or 'does not reject' a null hypothesis. The BF_{01} can provide evidence *for* the null hypothesis (if $BF_{01}>1$) as well as evidence for the alternative. A significance test may 'not reject' a null hypothesis but cannot provide evidence that supports the null.

The posterior probability of QI is proposed as a diagnostic for summarizing the evidence about the presence of QI. Transforming the posterior probability to the Bayes factor, BF_{01}, facilitates the interpretation by comparing the prior odds in favor of QI to the posterior odds of QI. A new test is proposed in which the null hypothesis of no QI is rejected when BF_{01} is less than a bound m.

Similarly, a diagnostic for 'clinically meaningful QI' can be defined, for example, by specifying values c_b and c_h, $c_h<c_b$, to represent clinically meaningful differences: benefit and harm, respectively. The BF_{01} for the following null and alternative hypotheses can be used as a diagnostic:

H_{0*}: There exists no pair $j\neq j'$ such that $\phi_j>c_b$ and $\phi_{j'}<c_h$.
H_{1*}: There exists at least one pair $j\neq j'$ such that $\phi_j>c_b$ and $\phi_{j'}<c_h$.

An example is given in Section 2.2.

2.1. Motivating example with binary responses

This method was motivated by the design of a multi-center trial for cell transplantation. In the trial, 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 multi-center trial with seven centers and subjects randomized to either a transplant or a 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 QI.

Let y_{ij} denote the observed number of successes out of n_{ij} subjects in the ith 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 treatment in each center, y_{ij}, is a draw from a Binomial distribution

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

For the logit link, define $\theta_{ij}=\text{logit}(p_{ij})=\log(p_{ij}/(1-p_{ij}))$, 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,

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

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

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

Therefore, define $\mu=\mu_2-\mu_1$ and $\omega=\sigma\sqrt{2}$ in equation (2).

The prior distribution was constructed with the aid of existing data from the seven 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 $\alpha/\beta=\frac{4}{3}$. The population parameters (σ^2, μ_1, μ_2) are assumed independent and both μ_1 and μ_2 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.

This trial has been designed but not completed and the simulation studies in Section 3 are motivated by this example. Data from a second, completed, trial are used below to illustrate a Bayesian subgroup analysis.

2.2. IHAST example

Intraoperative Hypothermia for Aneurysm Surgery Trial (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

Table I. Intraoperative Hypothermia for Aneurysm Surgery Trial [22].

	% favor. outcome			Posterior	
Gender	Hypo.	Normo.	% diff (95% Conf. Int.)	Mean	Cred. Int.
Male	69% ($\frac{120}{174}$)	57% ($\frac{97}{171}$)	0.12 (0.02, 0.22)	0.12	(0.02, 0.22)
Female	64% ($\frac{209}{325}$)	66% ($\frac{217}{330}$)	−0.02 (−0.09, 0.06)	−0.01	(−0.09, 0.06)

an acute subarachnoid hemorrhage (SAH) undergoing an open craniotomy to clip their aneurysms [22]. The outcome is binary: a favorable outcome is defined as Glasgow outcome score (GOS) equal to 1, 90 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' [22]. A total of 1000 subjects were followed postoperatively and GOS evaluated on or about 90 days after surgery. Randomized treatment assignment was stratified by center (30 centers). 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 percent favorable outcome on hypothermia vs 63 percent favorable outcome on normothermia, odds ratio = 1.14, 95 percent confidence interval: 0.88–1.48).

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 percent of males ($\frac{120}{174}$) were classified as having a good outcome, as compared with 57 percent ($\frac{97}{171}$) in the normothermia group. Among women, 64 percent (209/325) in the hypothermia group had a GOS score of 1, as compared with 66 percent ($\frac{217}{330}$) in the normothermia group (Table I). These results suggest that perhaps males benefit from intraoperative hypothermia during surgery, whereas in females there is either no effect, or hypothermia is harmful. The estimates and the confidence intervals are almost identical to the posterior means and the credible intervals.

The null hypothesis of no interaction of any kind is rejected ($p = 0.03$ by the LRT), and so the G&S and P&G tests are applied. Neither of these tests detects a difference in the sign of the treatment effect between genders ($p > 0.20$).

At the design stage, there was an expectation that the good outcome rate would be approximately 70 percent overall, and based on previous data it is unlikely to be less than 30 percent and not more than 93 percent. Because hypothermia may lead to additional complications, such as post-operative infections, a clinically meaningful difference was prespecified as 65 percent good outcome in the normothermia group and 75 percent in the hypothermia group. It was also anticipated that the treatment effect in men would be similar to the treatment effect in women, but there was considerable uncertainity.

The prior distribution for both μ_1 and μ_2 was therefore centered at logit(0.70) = 0.85 with a standard deviation of 0.85, so that conditional on $\mu_i = 0.85$ for any $i = 1, 2$ a 95 percent credible interval for p_{ij} is 0.30–0.93 for $j = 1, \ldots, n_j$. The Gamma($\alpha = 4, \beta = 2$) distribution on σ^{-2} reflects considerable uncertainty and a coefficient of variation of 0.5. The posterior means and credible intervals for the simple effects in each subgroup are almost identical to the frequentist estimates. The interpretation of the evidence about QI using the BF_{01} is however different than the frequentist tests.

The prior probability of QI is 0.31 as calculated from Equation (3). The posterior probability of QI is estimated to be 0.63, which is larger than the prior probability. BF_{01} is 0.267, indicating 'substantial' evidence for QI with Jeffreys' scale and 'positive' evidence with Kass and Raftery's scales. The Bayes factor quantifies the evidence concisely and accurately. In addition, as described later in Section 2.3 an exact test of size 0.05 using the Bayes factor gives a p-value of 0.03, indicating evidence for the presence of QI. As shown below, however, the interaction is unlikely to be clinically meaningful. Several additional prior distributions were also used, including some where the prior mean for μ_1 was logit(0.65) and that for μ_2 was logit(0.75). In all cases, very similar results were obtained for both posterior probability of QI and the BF_{01}.

A pre-specified clinically meaningful difference in IHAST is 65 vs 75 percent good outcomes for normothermia vs hypothermia, respectively. This corresponds to a difference of $c_b = 0.48$ on the log odds scale. Because hypothermia involves additional complications, as well as additional cost, the definition of harm is therefore $c_h = 0$. A clinically meaningful QI by gender can therefore be defined as occurring if $\phi_j > c_b = 0.48$ and $\phi_{j'} < c_h = 0$ for some $j \neq j'$. Joint prior probabilities for treatment effect of females and males lying in different regions of $\mathbb{R}^2$ are given in Table II using the prior distribution $\mu_i \sim \text{Normal}(0.85, 0.85^2)$ and $\sigma^2 \sim$ Inv-Gamma($\alpha = 4, \beta = 2$). For example, the joint prior probability of both treatments being harmful (ϕ_1 and ϕ_2 less than c_h) is 0.33, and the prior probability of both being clinically beneficial (both larger than c_b) is 0.23. The prior probability of clinically meaningful QI is 0.24 (corresponding to 0.12 + 0.12).

The corresponding joint posterior probabilities are also given in Table II. The most dramatic increase from prior to posterior probabilities is concentrated on the combination of two regions: one where the effect of hypothermia is not clinically meaningful for either males or females (prior to posterior probability 0.01 to 0.16) and the second where the effect of hypothermia is clinically beneficial for males, and for females the effect is neither clinically beneficial or harmful (prior to posterior probability of QI is 0.05–0.52). The possibility that hypothermia is harmful for males has been essentially ruled out (posterior probability 0.03), but the possibility that hypothermia is harmful for females has not been ruled out (posterior probability 0.32), although this probability is lower than the prior probability (the prior probability is 0.50 and the corresponding $BF_{01} = 2.125$, supporting the null hypothesis of no clinically meaningful QI).

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

	Males		
Females	$\phi_1 < c_h$	$c_h < \phi_1 < c_b$	$\phi_1 > c_b$
$\phi_2 < c_h$	0.33 (0.01)	0.05 (0.09)	0.12 (0.22)
$c_h < \phi_2 < c_b$	0.05 (0.02)	0.01 (0.16)	0.05 (0.52)
$\phi_2 > c_b$	0.12 (0.00)	0.05 (0.00)	0.23 (0.00)

The posterior probabilities are in parentheses, after the corresponding prior probabilities.

2.3. Bayes test with exact size, IHAST example

Define $\underline{\phi} = (\phi_1, \ldots, \phi_N)^T$ to be the vector of treatment effects in the N subgroups. A BF_{01} test with exact Type I error probability can be constructed, where $\underline{\phi}$ is not fixed, but has a distribution over the region of the parameter space corresponding to H_0. This distribution corresponds to the prior distribution restricted to this subspace. A large number of sets of values of θ_{ij} can be generated from the prior distribution. Each set of θ_{ij} corresponds either to H_0 or to H_1 defined in (1). For the θ_{ij} that correspond to H_0, data y_{ij} are generated and the BF_{01} calculated. In this way, a distribution of BF_{01} under the null hypothesis of no QI is generated, reflecting the variability over the space of corresponding θ_{ij}. A critical value is defined as the $\alpha \times 100$th percentile of the distribution to give an exact test of type I error probability α. The corresponding power can also be estimated by generating a data set for each of the θ_{ij} corresponding to H_1 and calculating the BF_{01}. The power is estimated by the proportion of these BF_{01} corresponding to H_1 less than the critical value.

IHAST data for gender are used as an example. The prior distribution is assumed to be that from the IHAST example in Section 2.2: the between-center variance has an inverse gamma distribution with parameters $\alpha = 4$ and $\beta = 2$ and both μ_1 and μ_2 have a normal distribution with mean and standard deviation of 0.85, and both hyper-parameters are independent. A total of 10^5 sets of values of θ_{ij}, for $i = 1, 2$ and $j = 1, 2$, are generated from this prior distribution to give 10^5 sets $\underline{\phi} = (\phi_1, \phi_2)^T$. QI occurs when there is at least one negative ϕ_j and at least one positive $\phi_{j'}$: a proportion of the 10^5 values of $\underline{\phi}$ are expected to correspond to this alternative hypothesis, the proportion is given by the expression in equation (3).

For each of the 10^5 sets of θ_{ij}, data are generated and the posterior probability of QI and BF_{01} is calculated. For the parameter values corresponding to $\underline{\phi}$ in the null hypothesis region, the values of BF_{01} that are less than or equal to the observed BF_{01} are calculated. The frequency of these as extreme or more extreme than the observed BF_{01} are reported as the exact p-value of the test: p-value $= 0.03$, providing evidence for the presence of QI. An exact test of clinically meaningful QI was also constructed, using the definition in Section 2.2, and the p-value is 0.25 and so there is no strong evidence that any such QI is clinically meaningful. These exact results are therefore consistent with the results from using the guidelines for the interpretation of BF_{01} in [21, p. 55]. Recall that neither the P&G nor G&S test was significant ($p > 0.2$), reflecting the low power of the frequentist tests for QI. The BF_{01} has indicated a suggestion that there may be QI, but it is not clinically meaningful.

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 $n_. = \sum_{j=1}^{N} n_j$. The simulation studies are based on the motivating example of Section 2.1 with $N = 7$ centers and binary responses. BF_{01} for each simulated data set is calculated using the prior distribution, where 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 with $\alpha = 2$ and $\beta = 1.5$. The prior probability of QI, given in (3), for $N = 7$ centers is estimated by Monte Carlo to be 0.37. This prior probability depends only on the prior distribution and the number of centers. The ability of BF_{01} to detect QI is examined and compared with the power of the methods of P&G and G&S. For each combination of $\underline{\phi}$ 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 tests that compare BF_{01} with 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 with each other. In practice, the exact value of the BF_{01} should be reported to quantify the evidence for QI, 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 as is the impact of balance, which was a particular concern among the investigators in our motivating example.

Because the data are simulated from a model with the p_{ij} fixed, these are power and size calculations in the frequentist framework for fixed p_{ij}: and the power of the BF_{01} test is the frequentist power of using BF_{01} as a test statistic. Note that for each specified value of ϕ, corresponding success probabilities p_{ij}, are required: in the simulations, 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 $p_2 = 0.731$.

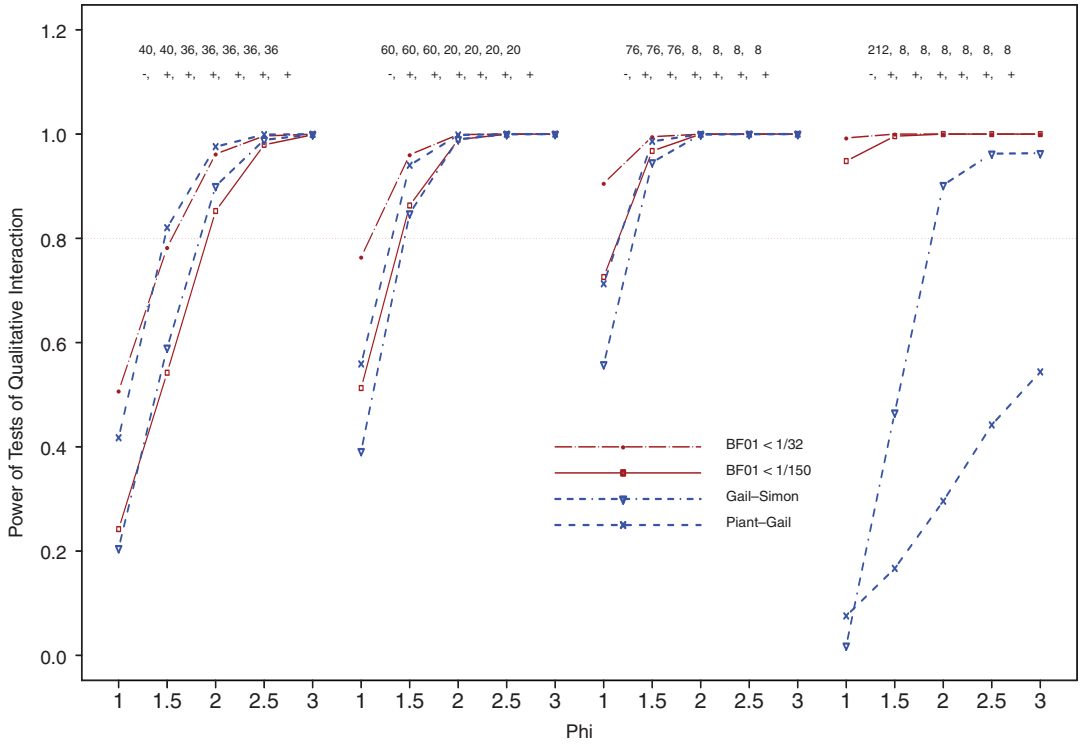

Figure 1. Power of QI tests where log odds treatment effect is negative in the first center and positive in the remainder: $\underline{\phi}=(-\phi,\phi,\phi,\phi,\phi,\phi,\phi)$ for $\phi=1$, 1.5, 2, 2.5, 3 and four designs with $n_{\cdot}=260$ ranging from most balanced $\mathbf{n}=(40,40,36,36,36,36,36)$ to most unbalanced $\mathbf{n}=(212,8,8,8,8,8,8)$.

For each of 5000 simulated data sets at each combination of $\underline{\phi}$ and **n** fixed parameter values, the LRT is used to test for any interaction by fitting two fixed effects GLMs, one with main effects only and the second with interaction, and comparing the difference in deviance with 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 QI test and the LRT are significant at 0.05. Both the protected and the unprotected tests were examined in the simulations and were almost identical. Only the protected results are reported. The posterior probability of QI is calculated irrespectively of the result of the LRT. The data sets are generated in R [23], and G&S and P&G tests are applied in R. The data set is passed to WinBUGS [24] using R2WinBUGS [25] for the posterior calculations and the results returned to R to calculate BF_{01}.

3.2. 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 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. Figure 1 gives the results for a total sample size of 260 and four designs ranging from most balanced to extremely unbalanced. Figure 1 shows that when the G&S and P&G tests have low power (in the most unbalanced cases and with smaller sample sizes), the BF_{01} tests have high power. When the G&S and P&G tests have power over 80 percent (in the more balanced cases), the BF_{01} tests have similar power.

Figure 2 shows the impact of increasing the sample size from $n_{\cdot}=130$ to $n_{\cdot}=520$ in the balanced case with $\underline{\phi}=(-\phi,\phi,\phi,\phi,\phi,\phi,\phi)$. The P&G test is the most powerful for $n_{\cdot}=130$ and the $BF_{01}<\frac{1}{32}$ is the most powerful for $n_{\cdot}=520$.

3.3. Power when treatment is harmful in three centers

Figure 3 provides the results when the treatment is beneficial in four centers and harmful in three. The pattern is similar to Figure 1, but with higher power. Additional simulations give similar results [26].

The P&G test is powerful for balanced designs, but as balance declines, the power of this test decreases dramatically. P&G [9] 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.

3.4. Type I error rate

To examine the Type I error probability, data are generated under several settings corresponding to the absence of QI in examples similar to those for the power calculations. It was assumed that ϕ_j is constant for $j=1,\ldots,7$. A total of seven different sets of

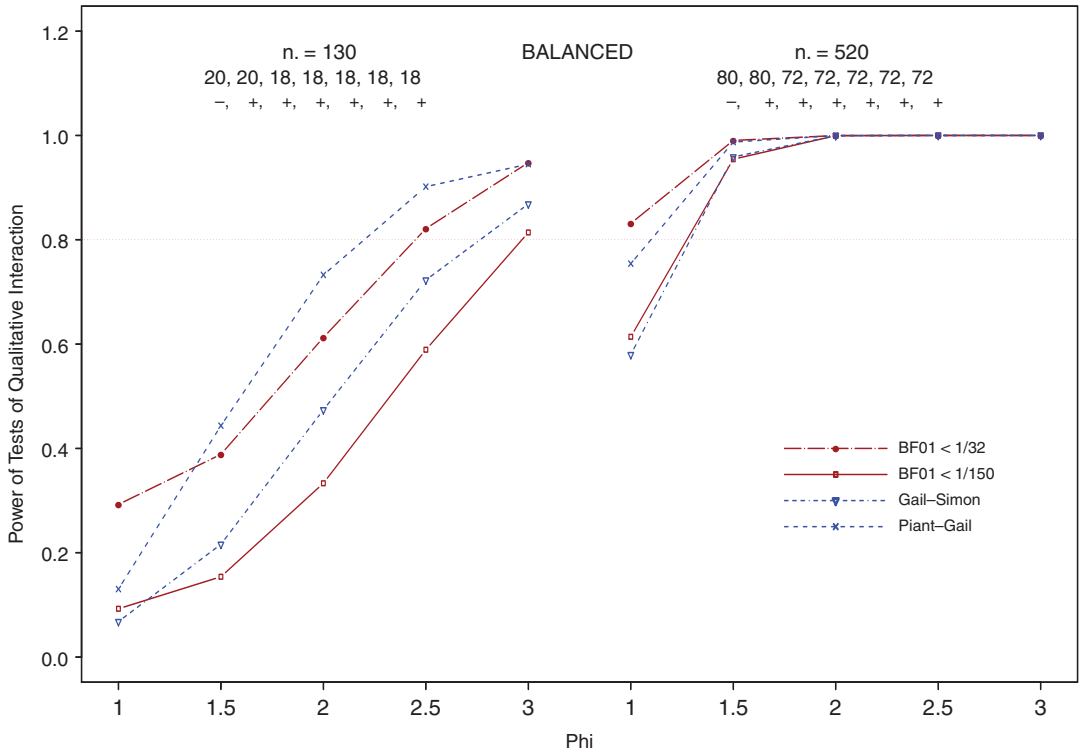

Figure 2. Power of QI tests for two balanced designs: $\mathbf{n}=(20,20,18,18,18,18,18)$ and $\mathbf{n}=(80,80,72,72,72,72,72)$. The total sample size is either $n_{\cdot}=130$ or $n_{\cdot}=520$. The subgroup treatment effect is negative in the first center, and positive in the remaining centers: $\underline{\phi}=(-\phi,\phi,\phi,\phi,\phi,\phi,\phi)$ for $\phi=1,\ 1.5,\ 2,\ 2.5,\ 3$.

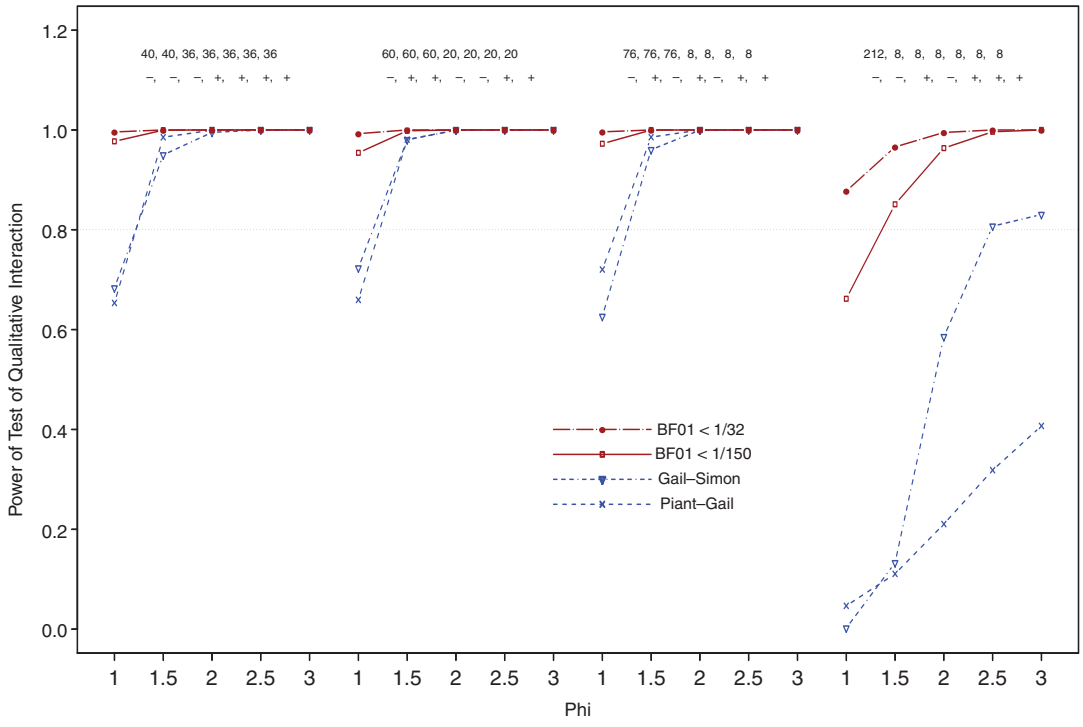

Figure 3. Power of QI tests where $\underline{\phi}$ is such that $|\phi_j|=\phi$ for $j=1,\ldots,7$, and $\phi=1,\ 1.5,\ 2,\ 2.5,\ 3$. The log odds treatment effect is negative in three centers and positive in the remainder as indicated. The total sample size is $n_{\cdot}=260$ in all cases.

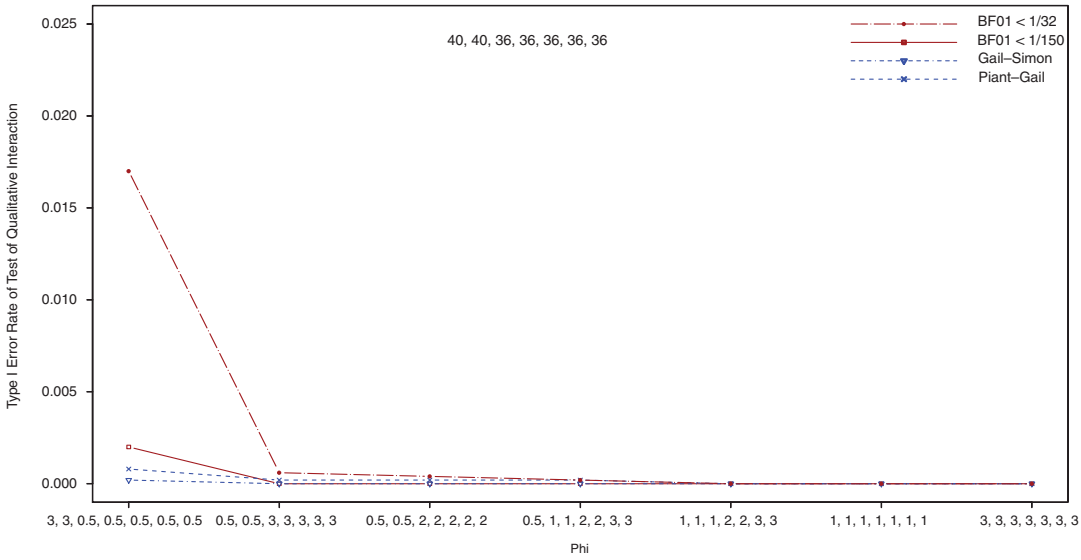

Figure 4. Type I error probability under seven different values of $\underline{\phi}$, where $\underline{\phi}$ satisfies the no QI hypothesis, for the design $\mathbf{n}=(40, 40, 36, 36, 36, 36, 36)$.

values of $\underline{\phi}$ were chosen for the the null hypothesis H_0 in (1). The total sample size was assumed to be either 65, 130, or 260. Figure 4 gives the results 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 4 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 percent 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 of these cases, the Type I error rate of the Bayesian test was very large: close to 50 percent. If there is at least one j' such that $\phi_{j'}=0$ and all other ϕ_j are of the same sign, $j=1,\ldots,N$, the parameter $\underline{\phi}$ is on the boundary of the region in the parameter space corresponding to H_0 in (1). Assuming that the posterior mean for $\underline{\phi}$ converges to the true value and that the posterior distribution is approximately normal, then the posterior probability of QI, when at least one ϕ_j is zero, will be approximately 0.50 and the posterior odds of QI are approximately one. It should therefore be expected that the Type I error rate is large in this case. The increased power of comparing the BF_{01} with a guideline is potentially at the expense of increased size. In contrast, the size of the P&G and G&S tests are very conservative in order to control type I error over the null hypothesis, at the expense of very low power at many parameter values. The exact test based on the BF_{01}, described in Section 2.3, addresses this by controlling the exact average Type I error, averaged over the space corresponding to the null hypothesis, weighted proportional to the prior distribution constrained to the parameter space under the null hypothesis. Examination for the presence of clinically meaningful QI adds to the interpretation of the subgroup analysis.

4. Conclusions and discussion

This paper has developed an approach to summarize the evidence for QI using the Bayes factor, where QI can be defined from the context, for example, as clinically meaningful QI, as in the IHAST example. The Bayesian approach requires the specification of a prior distribution. The prior distributions used here assume that the treatment responses in each subgroup are exchangeable: 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. The G&S and P&G statistics are based on a fixed effect model and do not incorporate any prior beliefs.

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.

Section 3 examined the use of the BF_{01} as a test statistic. A test of significance is designed, typically, to look for evidence *against* a null hypothesis in favor of an alternative. The BF_{01}, in contrast, concisely summarizes the evidence *for and against* the null hypothesis of no QI. For example, a large *p*-value or lack of significance does not provide evidence *for* the null but, in contrast, a large BF_{01}, greater than one, provides evidence in support of the null.

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 for situations where all the treatment effects are of the same sign, but there is at least one

that is equal or close to zero. When there is no QI but the treatment effect in one subgroup is exactly zero, the parameter values are on the boundary of the parameter space corresponding to the null hypothesis of no QI. Alternatively, when the treatment effects, in all subgroups are identical, and also large in absolute value, all four tests are highly conservative. The inflated Type I error probability can be addressed through constructing an exact test based on the BF_{01}, where the region of the null hypothesis is considered and the Type I error rate is averaged over the space.

In the IHAST example of Section 2, the BF_{01} provided strong evidence for QI, but little evidence for clinically meaningful QI. In contrast, the G&S and P&G tests provided no evidence for QI ($p>0.2$). The conclusion of the investigators that hypothermia should not be routinely recommended remains valid.

Irrespective of the simulations in Section 3 and irrespective of the frequentist significance testing, from a Bayesian perspective the BF_{01} concisely summarizes evidence for and against QI and should be considered as a diagnostic measure.

Acknowledgements

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. We are also very grateful to the referees and editor for their suggestions that have greatly improved this paper.

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. DOI: 10.1002/sim.1295.
4. Follmann D. Subgroups and interactions. In *Advances in Clinical Trials Biostatistics*, Geller NL (ed.). CRC: Boca Raton, 2003; 124–142.
5. Peto R. Statistical aspects of cancer trials. In *Treatment of Cancer*, Halnan KE (ed.). Chapman & Hall: London, 1982; 867–871.
6. Gail M, Simon R. Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics* 1985; **41**:361–372.
7. Silvapulle MJ. Tests against qualitative interaction: exact critical values and robust tests. *Biometrics* 2001; **57**:1157–1165.
8. Russek-Cohen E, Simon RM. Qualitative interactions in multi-factor studies. *Biometrics* 1993; **49**:467–477.
9. Piantadosi S, Gail MH. A comparison of the power of two tests for qualitative interactions. *Statistics in Medicine* 1993; **12**:1239–1248.
10. Li J, Chan ISF. Detecting qualitative interactions in clinical trials: an extension of range test. *Journal of Biopharmaceutical Statistics* 2006; **16**:831–841. DOI: 10.1080/10543400600801588.
11. Pan G, Wolfe D. Test for qualitative interaction of clinical significance. *Statistics in Medicine* 1997; **16**:1645–1652.
12. Yan X, Su X. Testing for qualitative interaction. *Encyclopedia of Biopharmaceutical Statistics* 2005; **1**:1–8. DOI: 10.1081/E-EBS-120040186.
13. 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.
14. Boos DD, Brownie C, Zhang J. Estimating the magnitude of interaction. *Institute of Statistics Mimeo Series, 2285*, North Carolina State University, Raleigh, NC, 1996. Available from: http://www4.stat.ncsu.edu/ boos/papers.html, accessed on 17/08/2009.
15. Duncan DB. A Bayesian approach to multiple comparisons. *Technometrics* 1965; **7**(2):171–222.
16. Dixon DO, Simon R. Bayesian subset analysis. *Biometrics* 1991; **47**:871–881.
17. 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. DOI: 10.1002/sim.2420.
18. Good IJ. *Good Thinking*: *The Foundations of Probability and its Applications*. University of Minnesota Press: Minneapolis, 1983; 36.
19. Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995; **90**(430):773–795.
20. Jeffreys H. *Theory of Probability* (3rd edn). Oxford University Press: Oxford, 1961; 432.
21. Spiegelhalter DJ, Abrams KR, Myles KR. *Bayesian Approaches to Clinical Trials and Health-care Evaluation*. Wiley: England, 2007; 55.
22. 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.
23. R Development Core Team. *R*: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. Available from: http://www.R-project.org. ISBN: 3-900051-07-02005. accessed 17/08/2009.
24. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS—a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing* 2000; **10**:325–337.
25. Sturtz S, Ligges U, Gelman A. R2WinBUGS: a package for running WinBUGS from R. *Journal of Statistical Software* 2005; **12**(3):1–16.
26. Bayman EO. Bayesian hierarchical models for multi-center clinical trials: power and subgroup analysis. *Ph.D. Dissertation*, Biostatistics, The University of Iowa, Iowa City, IA, 2008.