STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:1495–1508
Published online 10 February 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2052

Statistical design and analysis of pharmacogenetic trials

Patrick J. Kelly[1,*,†], Nigel Stallard[1] and John C. Whittaker[2]

[1] *Medical and Pharmaceutical Statistics Research Unit, The University of Reading, Reading RG6 6FN, U.K.*
[2] *Department of Epidemiology and Public Health, Imperial College, St Mary's Campus, Norfolk Place, London W2 1PG, U.K.*

SUMMARY

Pharmacogenetic trials investigate the effect of genotype on treatment response. When there are two or more treatment groups and two or more genetic groups, investigation of gene–treatment interactions is of key interest. However, calculation of the power to detect such interactions is complicated because this depends not only on the treatment effect size within each genetic group, but also on the number of genetic groups, the size of each genetic group, and the type of genetic effect that is both present and tested for. The scale chosen to measure the magnitude of an interaction can also be problematic, especially for the binary case.

Elston *et al.* proposed a test for detecting the presence of gene–treatment interactions for binary responses, and gave appropriate power calculations. This paper shows how the same approach can also be used for normally distributed responses. We also propose a method for analysing and performing sample size calculations based on a generalized linear model (GLM) approach. The power of the Elston *et al.* and GLM approaches are compared for the binary and normal case using several illustrative examples. While more sensitive to errors in model specification than the Elston *et al.* approach, the GLM approach is much more flexible and in many cases more powerful. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: pharmacogenetics; sample size; power; clinical trials

1. INTRODUCTION

The purpose of a pharmacogenetic trial is to investigate a gene or genes that are responsible for interpatient variation in drug responses. An important aspect of the design of pharmacogenetic trials, as of any clinical trial, is the determination of the sample size required to detect with specified power a clinically significant difference between groups. In a standard clinical trial there are usually two or more treatment groups and the purpose is to determine

[*] Correspondence to: Patrick J. Kelly, MPS Research Unit, The University of Reading, PO Box 240, Earley Gate, Reading RG6 6FN, U.K.
[†] E-mail: patrick.kelly@reading.ac.uk

Contract/grant sponsor: Variagenics Inc

Received February 2003
Accepted May 2004

Copyright © 2005 John Wiley & Sons, Ltd.

whether there is a difference between treatments, that is, a treatment effect. In this paper we consider pharmacogenetic study designs where there are two or more treatment groups and two or more genetic groups. In this case, two additional questions are of interest: firstly, whether there is a difference between genetic groups, that is, a gene effect; and secondly, whether different genetic groups respond differently to treatment, that is, whether there is a gene–treatment interaction effect. The calculation of power to detect a main treatment effect is straightforward, but determination of the power to detect a gene or interaction effect is slightly more complicated since this depends not only on treatment effect size within each genetic group but also on the number of genetic groups, the genetic group frequencies and the type of genetic effect that is both tested for and present in the data [1]. Even what is meant by an interaction may be problematic, since this depends on the scale of measurement [2]. This is particularly true for binary data; the main effects of gene and treatment with no interaction on one scale, for example, the logit scale, may correspond to an interaction on a different measurement scale, such as the probability difference.

In this paper, we propose a method for analysing pharmacogenetic studies based on generalized linear models (GLM). The approach is therefore appropriate for responses from any exponential family distribution, but we shall concentrate on binary and Gaussian responses. Power calculations for binary and Gaussian responses will also be developed for the GLM approach.

An alternative method for the analysis of, and power calculations for, pharmacogenetic trials with binary responses has been proposed by Elston *et al.* [3]. We show how this approach can be easily adapted to the case of Gaussian responses and compare it with the method based on generalized linear models in this and the binary data case.

This paper is divided into six sections. Section 2 introduces the notation that will be used throughout this article. Section 3 outlines the method by Elston *et al.* and explains how it may be applied in the case of Gaussian data. Section 4 describes the method based on GLM for both Gaussian and binary data. Section 5 compares the two methods using illustrative examples. Section 6 is a discussion, which includes a summary of the advantages and disadvantages of the two approaches.

2. NOTATION

Suppose we have $I+1$ genetic groups and $J+1$ treatment groups, so that the total number of gene-by-treatment groups is $(I+1)(J+1)$. For example, in the simplest genetic scenario we would consider a single diallelic locus, say with alleles A and a, which gives three genetic groups corresponding to the genotypes aa, Aa and AA, respectively, so that $I=2$.

Let n_{ij} denote the number of subjects in the ith genetic and jth treatment group for $i=0,\ldots,I,\ j=0,\ldots,J$ and n, $n_{i.}$ and $n_{.j}$ denote, respectively, the total number of subjects, the number of subjects in the ith genetic group and the number of subjects in the jth treatment group, given by $n=\sum_{i=0}^{I}\sum_{j=0}^{J}n_{ij}$, $n_{i.}=\sum_{j=0}^{J}n_{ij}$ and $n_{.j}=\sum_{i=0}^{I}n_{ij}$.

Let Y_{ijk} denote the response of the kth individual in the ijth group. If the clinical response is a quantitative measure then we shall assume that the response has a normal distribution, where the mean can be different between subgroups but the within-group variance is constant, that is, $Y_{ijk}\sim \mathrm{N}(\mu_{ij},\sigma^2)$ for $i=0,\ldots,I,\ j=0,\ldots,J,\ k=1,\ldots,n_{ij}$. We shall use $\overline{y}_{ij}$ to denote the

Copyright © 2005 John Wiley & Sons, Ltd.

observed group mean. If the clinical response is binary then the probabilities of success and failure is given by $P(Y_{ijk}=1)=\pi_{ij}$ and $P(Y_{ijk}=0)=1-\pi_{ij}$, respectively, with s_{ij} indicating the number of observed successes and $p_{ij}=s_{ij}/n_{ij}$ denoting the observed response proportion.

3. ELSTON *ET AL.* APPROACH

Elston *et al.* [3] consider binary data in study designs with two treatment groups or one treatment at several doses (dose–response study). Their approach is based on the idea of testing linear combinations of the success probabilities in the different gene–treatment groups, and assumes we wish to test for gene or interaction effects at a diallelic locus. The general null hypothesis is thus

$$H_0 : \sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}\pi_{ij}=0$$

for a given set of coefficients w_{ij}, such that $\sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}=0$. The test statistic proposed to test this null hypothesis is

$$t=\frac{\sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}\,p_{ij}}{\sqrt{\sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}^{2}\hat{V}(p_{ij})}}, \tag{1}$$

where $\hat{V}(p_{ij})$ is the estimated variance of p_{ij}. Elston *et al.* consider two different variance estimates of $\hat{V}(p_{ij})$:

(a) $\hat{V}(p_{ij})=p_{ij}(1-p_{ij})/n_{ij}$; and
(b) $\hat{V}(p_{ij})=p_j(1-p_j)/n_{ij}$, where $p_j=\sum_i s_{ij}/n_{.j}$.

The variance estimate (a) is consistent for any values of π_{ij}, while estimate (b) is consistent under any hypothesis in which the π_{ij} do not depend on i (genotype) [3].

There are several potential alternative hypotheses of interest in a pharmacogenetic trial. The type of alternative hypothesis to be tested is determined by w_{ij}. Elston *et al.* consider examples for study designs with two parallel treatment groups ($J=1$) and illustrates two types of hypothesis tests: testing for an allele effect and testing for an allele–treatment interaction effect. Three different allele effects are of particular interest: additive, dominant and recessive. Similarly, there are three allele–treatment interaction effects of potential interest. Tables I and II show the coefficients, w_{ij}, that can be used in order to test for these allele and interaction effects, respectively.

Table I. Coefficients for testing different allele effects.

Genotype	Coefficients	Additive	Dominant	Recessive
aa	$w_{00}=w_{01}$	−1	−2	−1
Aa	$w_{10}=w_{11}$	0	1	−1
AA	$w_{20}=w_{21}$	1	1	2

Copyright © 2005 John Wiley & Sons, Ltd.

Table II. Coefficients for testing different allele–treatment interaction effects.

Genotype	Coefficients		Additive		Dominant		Recessive	
aa	w_{00}	w_{01}	1	−1	2	−2	1	−1
Aa	w_{10}	w_{11}	0	0	−1	1	1	−1
AA	w_{20}	w_{21}	−1	1	−1	1	−2	2

For large samples we can use the normal approximation to the binomial distribution to say that the test statistic given by (1) is approximately normally distributed with unit variance and mean

$$E_A = \frac{\sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}\pi_{ij}}{\sqrt{\sum_{i=0}^{2}\sum_{j=0}^{J}(w_{ij}^2\pi_{ij}(1-\pi_{ij})/n_{ij})}}, \tag{2}$$

which is zero under the null hypothesis.

If $g_i = n_{i.}/n$ and $t_j = n_{.j}/n$ denote the observed relative frequencies of the ith genetic group and the jth treatment group, respectively, (2) can be rewritten as

$$E_A = \frac{\sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}\pi_{ij}}{\sqrt{\frac{1}{n}\sum_{i=0}^{2}\sum_{j=0}^{J}(w_{ij}^2\pi_{ij}(1-\pi_{ij})/g_it_j)}}.$$

By standard theory [4], the power to reject the null hypothesis, based on this test statistic for a two-sided test of size α, is approximately equal to

$$1-\beta = \Phi(E_A - z_{\alpha/2}) + \Phi(-E_A - z_{\alpha/2}) \tag{3}$$

where Φ is the cumulative standard normal distribution and z_u denotes the $(1-u)$th percentile of the standard normal distribution. By substituting (2) into (3) and solving for n, the sample size required to reject the null hypothesis with power $(1-\beta)$ is approximately

$$n = \left(\frac{z_{\alpha/2}+z_\beta}{\sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}\pi_{ij}}\right)^2 \sum_{i=0}^{2}\sum_{j=0}^{J}\left[\frac{w_{ij}^2\pi_{ij}(1-\pi_{ij})}{g_it_j}\right].$$

For samples of realistic size, the normal approximation for the test statistic given by (1) may be poor. This may be of particular concern when tests are conducted for extremely small α, due to correction for multiple testing, with the result that the normality assumption is used for the extreme tails of the distribution. Additionally, the mean and variance of the test statistic are calculated assuming that the relative frequency of genotype groups are fixed and known. In practice, n_{ij} is a random variable since we are unlikely to know the subjects' genotype before the start of the study. We can produce approximate power calculations by replacing g_i for $i = 0,\ldots,I$ in (2) with their expected values. For example, if we assume Hardy–Weinberg equilibrium (HWE), the expected genotype frequencies are $E(g_0) = (1-q)^2$, $E(g_1) = 2q(1-q)$ and $E(g_2) = q^2$, where q is the expected frequency of the A allele. However, this approximation may be poor if either the total sample size is small or the relative expected frequency of at least one genotype group is small.

Elston *et al.* suggest that these problems in small samples can be overcome by using simulations to estimate the exact distributions of the test statistic under the null and alternative hypotheses. Under the alternative hypothesis within each treatment group, the number of responders and non-responders in each genetic group are simulated from a six-nomial distribution with probabilities $E(g_0)\pi_{0j}$, $E(g_0)(1-\pi_{0j})$, $E(g_1)\pi_{1j}$, $E(g_1)(1-\pi_{1j})$, $E(g_2)\pi_{2j}$, and $E(g_2)(1-\pi_{2j})$. For testing an allele effect or an allele–treatment interaction effect Elston *et al.* simulate data under the same null hypothesis, which assumes the proportion of responders within each treatment group do not depend on i, so that the data are simulated from a binomial distribution with parameters $t_j n$ and $\sum_{i=0}^{2} E(g_i)\pi_{ij}$. However, it should be noted that this null hypothesis is incorrect if testing for an interaction effect when a main allele effect is also present. This is discussed further in Section 6.

3.1. Modification for normally distributed data

Although Elston *et al.* proposed their method for trials with binary responses, in the large sample case it is based on the asymptotic normality of the test statistic given by (1). Adapting the method to the case of Gaussian data is therefore straightforward.

For pharmacogenetic trials with normally distributed data the parameters of interest are the means of each gene–treatment group. The general null hypothesis is of the form

$$H_0 : \sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}\mu_{ij} = 0$$

The corresponding test statistic

$$t = \frac{\sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}\bar{y}_{ij}}{\sqrt{\sigma^2 \sum_{i=0}^{2}\sum_{j=0}^{J}(w_{ij}^2/n_{ij})}} \tag{4}$$

is normally distributed if the within-group variance, σ^2, is known, or is approximately so for a sufficiently large sample if σ^2 is unknown, with unit variance and mean

$$E_A = \frac{\sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}\mu_{ij}}{\sqrt{\sigma^2 \sum_{i=0}^{2}\sum_{j=0}^{J}(w_{ij}^2/n_{ij})}} \tag{5}$$

Under the null hypothesis, $E_A = 0$. The power to reject the null hypothesis at the two-sided level α is given by (3) where E_A is given by (5). The sample size required to give power $(1-\beta)$ at the specified alternative hypothesis is

$$n = \left(\frac{z_{\alpha/2} + z_\beta}{\sum_{i=0}^{2}\sum_{j=0}^{J} w_{ij}\mu_{ij}}\right)^2 \sigma^2 \sum_{i=0}^{2}\sum_{j=0}^{J}\left[\frac{w_{ij}^2}{g_i t_j}\right]$$

Again, prospective power calculations may be performed by replacing the g_i by their expected values. In small samples, or if the frequency of one of more genotype groups is small, these may give a poor approximation, so that a simulation approach similar to that described by Elston *et al.* may be preferred.

Copyright © 2005 John Wiley & Sons, Ltd.

4. GENERALIZED LINEAR MODEL APPROACH

In this section we describe a GLM approach to the analysis of pharmacogenetic studies. The approach can be applied to responses from any exponential family distribution, but we shall concentrate on normally distributed or binary responses. The relevant test statistics and corresponding power calculations are outlined below.

We will model the mean response for the kth subject ($k=1,\ldots,n$), $\mu_k=E(Y_k)$, by

$$g(\mu)=\alpha+\beta'_G G_k+\beta'_T T_k+\beta'_{GT} GT_k$$

where $g(.)$ is a link function, β_G, β_T and β_{GT} are vectors of gene effects, treatment effects and interaction effects, respectively, and T_k, G_k and GT_k are corresponding vectors of variables, coded so as to ensure identifiability. To illustrate the approach, we shall consider the simplest case of two treatment groups and a single diallelic locus, giving three genetic groups. We would have a single treatment variable, such as

$$T_k=\begin{cases}-1 & \text{if } j=0 \text{ for } k\text{th subject}\\ 0 & \text{if } j=1 \text{ for } k\text{th subject}\end{cases}$$

A number of possible parameterizations can be used for the genetic variables, corresponding to different genetic models. For example, a single variable G_k defined by

$$G_k=\begin{cases}-1 & \text{if } i=0 \text{ for } k\text{th subject}\\ 0 & \text{if } i=1 \text{ for } k\text{th subject}\\ 1 & \text{if } i=2 \text{ for } k\text{th subject}\end{cases}$$

$$G_k=\begin{cases}-2 & \text{if } i=0 \text{ for } k\text{th subject}\\ 1 & \text{if } i=1 \text{ for } k\text{th subject}\\ 1 & \text{if } i=2 \text{ for } k\text{th subject}\end{cases}$$

or

$$G_k=\begin{cases}-1 & \text{if } i=0 \text{ for } k\text{th subject}\\ -1 & \text{if } i=1 \text{ for } k\text{th subject}\\ 2 & \text{if } i=2 \text{ for } k\text{th subject}\end{cases}$$

gives an additive, dominant or recessive model, respectively, whilst using the vector (G_{1k},G_{2k}), where

$$G_{1k}=\begin{cases}-1 & \text{if } i=0 \text{ for } k\text{th subject}\\ 0 & \text{if } i=1 \text{ for } k\text{th subject}\\ 1 & \text{if } i=2 \text{ for } k\text{th subject}\end{cases}$$

and

$$G_{1k}=\begin{cases}-1 & \text{if } i=0 \text{ for } k\text{th subject}\\ 2 & \text{if } i=1 \text{ for } k\text{th subject}\\ -1 & \text{if } i=2 \text{ for } k\text{th subject}\end{cases}$$

gives a general model in which no relationship is assumed between the three genotypic effects.

We shall assume, for the purposes of this paper, that the genetic model for the interaction effect is of the same form as that for the main genetic effect. For example, if the model

Copyright © 2005 John Wiley & Sons, Ltd.

includes an additive allele effect, the interaction effect, if it is included, is also additive. Thus, β_G and β_{GT} are each scalars for a specified genetic model (additive, dominant or recessive) and are bivariate vectors for the general model. Consequently, there are four model degrees of freedom for a specified genetic model (one for each of intercept, gene, treatment and interaction) and six for the general model (one for each of intercept and treatment, and two for each of gene and interaction).

Hypotheses can be tested by fitting restricted models in the usual GLM manner. In particular, we wish to test the null hypothesis that there is no interaction effect,

$$H_0 : \beta_{GT} = 0$$

which is a one degree of freedom test for a specified genetic model or a two degrees of freedom test for the general model. To allow direct comparison to Elston *et al.* we will mainly concentrate on specified genetic models.

4.1. Normally distributed data

When the data are normally distributed, the link function is the identity function, $g(\mu_k) = \mu_k$, and our model is a standard linear regression. Hypothesis tests can be conducted using an F test [5].

When the null hypothesis is true, F follows a central F distribution with degrees of freedom $v_1 = r - s$ and $v_2 = n - r$, where r and s are the number of parameters estimated in the model under the alternative and null hypotheses, respectively. When the null hypothesis is false and the fitted alternative model is correct then F has a non-central F distribution, F', with the same degrees of freedom as before plus a non-centrality parameter, λ_1, given by

$$\lambda_1 = \frac{\mu' X_1 (X_1' X_1)^{-1} X_1' \mu - \mu' X_0 (X_0' X_0)^{-1} X_0' \mu}{\sigma^2}$$

where μ is the $n \times 1$ vector of the true mean responses, μ_k, and X_1 and X_0 are the $n \times r$ and $n \times s$ matrices of explanatory variables under the alternative and null model, respectively.

When both the null and the alternative model are incorrect, for example, if an additive gene–treatment interaction is fitted when the true model has a recessive gene–treatment interaction, the F test statistic follows a doubly non-central F distribution, F''. The F'' has the same parameters as the non-central F distribution, plus another non-centrality parameter, λ_2, given by

$$\lambda_2 = \frac{n \sum_{i=0}^{I} \sum_{j=0}^{J} g_i t_j (\hat{Y}_{ij} - \mu_{ij})^2}{\sigma^2}$$

where $\hat{Y}_{jk}$ is the fitted value for the ith genetic and jth treatment group.

In practice, the F'' distribution is difficult to calculate and is usually approximated by the F' distribution [6]. Therefore, the power is usually approximated by

$$1 - \beta = \Pr(\mathrm{F}''_{v_1, v_2, \lambda_1, \lambda_2} > F_{v_1, v_2}(\alpha)) \approx \Pr\left(\mathrm{F}'_{v_1, v_2^*, \lambda_1} > \frac{v_2 + \lambda_2}{v_2} F_{v_1, v_2}(\alpha)\right)$$

where

$$v_2^* = \frac{(v_2 + \lambda_2)^2}{v_2 + 2\lambda_2}$$

Copyright © 2005 John Wiley & Sons, Ltd.

We can prospectively calculate power at significance level α, by assuming g_i fixed and equal to their expected values. The sample size required to achieve a specified power, $1-\beta$, can be determined by searching for the smallest value of n which satisfies $\Pr(\mathrm{F}''_{v_1,v_2,\lambda_1,\lambda_2} > F_{v_1,v_2}(\alpha)) \geqslant 1-\beta$.

4.2. Binary data

A common model for binary data is the logistic regression model, where $E(Y_i)=\pi_i$ and the link function is the logistic function, $g(\pi_i)=\pi_i/(1-\pi_i)$. Hypothesis tests for logistic regression models can be conducted using the likelihood ratio test (LRT) statistic [5], given by

$$D=2[l(\hat{\beta}_1)-l(\hat{\beta}_0)]$$

where $l(\hat{\beta}_0)$ and $l(\hat{\beta}_1)$ are the log-likelihood evaluated at the maximum likelihood estimates under the null and alternative model, respectively.

When the null hypothesis is true, D asymptotically has a central chi-square distribution with degrees of freedom equal to the difference in the number of parameters in the null and alternative hypotheses, $v_1=r-s$. When the null hypothesis is false D asymptotically has a non-central chi-square distribution, $\chi^2_{v_1,\lambda}$, with the same degrees of freedom as before and non-centrality parameter, λ, which is the expected value of D under the specified alternative hypothesis. Thus the power is given by

$$1-\beta=\Pr(\chi^2_{v_1,\lambda}>\chi^2_{v_1}(\alpha))$$

The sample size required to achieve a specified power, $1-\beta$, can be determined by searching for the smallest value of n that satisfies $\Pr(\chi^2_{v_1,\lambda}>\chi^2_{v_1}(\alpha))\geqslant 1-\beta$.

4.3. Power calculations based on simulations

As in Section 3, the asymptotic properties of the LRT statistic may not hold in small samples or for tests with very small α, and neither the LRT nor the F statistics may accurately follow the distributions given since these are based on the expected genotype frequencies. As an alternative, we can adopt the approach used by Elston *et al.* and estimate the distributions of the test statistic under the null and alternative hypotheses via simulation.

5. COMPARISON OF METHODS

In this section, the power of the test statistics described in Sections 3 and 4 are compared for pharmacogenetic trials which have two equally sized treatment groups and a single diallelic candidate locus, where the expected frequency of allele A is q.

This gives a total of six (3×2) gene–treatment groups. HWE will be assumed so that the expected relative frequencies of the three genotype groups are $(1-q)^2$, $2q(1-q)$ and q^2. The power of the test statistics in this section are calculated using their asymptotic distributions and assuming the genotype frequencies are fixed.

Copyright © 2005 John Wiley & Sons, Ltd.

Table III. The mean responses for the additive gene–drug interaction.

	Treatment groups			
Genotype groups	Placebo	Treatment	Overall	Genotype frequency
aa	0	0	0	$(1-q)^2$
Aa	0	$d/2$	$d/4$	$2q(1-q)$
AA	0	d	$d/2$	q^2
Overall	0	qd	$qd/2$	1

Table IV. The mean responses for the dominant gene–drug interaction.

	Treatment groups			
Genotype groups	Placebo	Treatment	Overall	Genotype frequency
aa	0	0	0	$(1-q)^2$
Aa	0	d	$d/2$	$2q(1-q)$
AA	0	d	$d/2$	q^2
Overall	0	$dq^2+2q(1-q)d$	$dq^2/2+q(1-q)d$	1

Table V. The mean responses for the recessive gene–drug interaction.

	Treatment groups			
Genotype groups	Placebo	Treatment	Overall	Genotype frequency
aa	0	0	0	$(1-q)^2$
Aa	0	0	0	$2q(1-q)$
AA	0	d	$d/2$	q^2
Overall	0	dq^2	$dq^2/2$	1

5.1. Quantitative traits

The power to detect a quantitative trait is calculated for three simple scenarios: when the true gene–drug interaction is additive (Table III), dominant (Table IV) and recessive (Table V). For the three scenarios, the mean response within the control group is zero for each genotype group. Within the treatment group, the mean response for the *aa* genotype is zero, the mean response for the *AA* genotype is d and the mean response for the *Aa* genotype is determined by the type of gene–drug interaction. A common variance, σ^2, is assumed within each genotype–treatment group. Figure 1 displays the power to detect an additive, dominant or recessive interaction effect using the Elston *et al.* test statistic shown in Section 3.1 and the F test, for the three scenarios, when $n=300$, $\alpha=0.05$, $d=1$, $\sigma^2=1$ and q takes a range of values between 0.1 and 0.9. We have also investigated a broad range of other scenarios, all of which give similar patterns of results.

Copyright © 2005 John Wiley & Sons, Ltd.

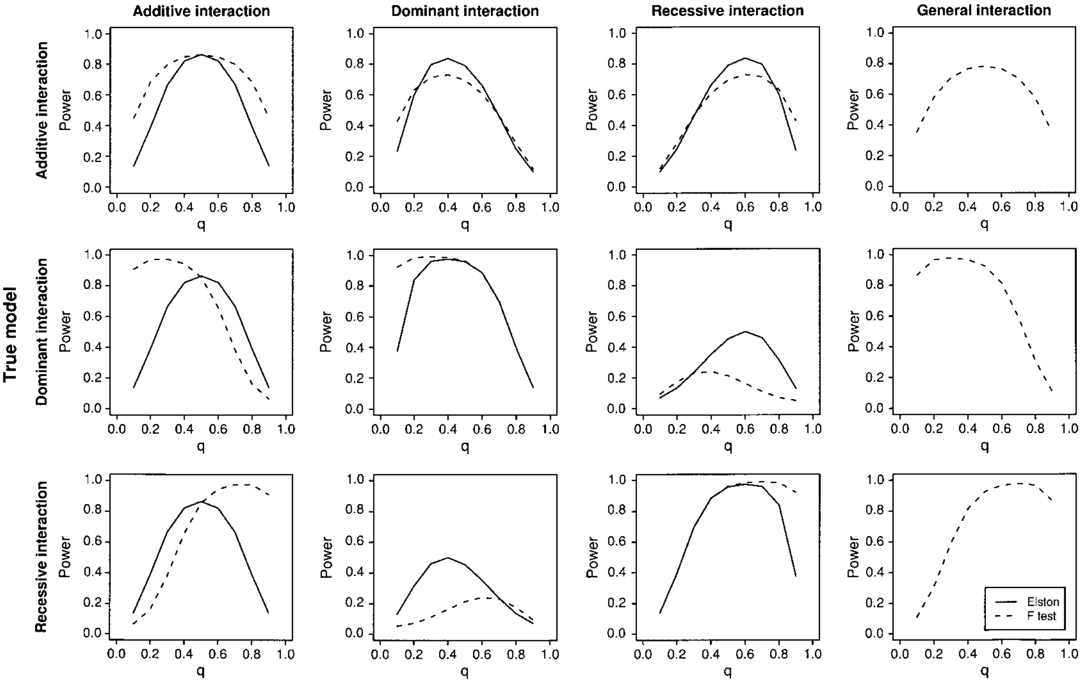

Figure 1. Power to detect interaction effects for normally distributed data using the Elston *et al.* and F test statistics, versus the allele frequency (q), for $n = 300$, $\alpha = 0.05$, $d = 1$, $\sigma^2 = 1$ and assuming HWE. The rows correspond to the true underlying models (see Tables III–V) and the columns correspond to the test for an interaction effect. For example, the diagram located at the first row and second column displays the power to detect a dominant interaction effect, when the true model has an additive interaction effect.

Figure 1 shows that the Elston *et al.* test statistic for testing a specific type of interaction effect is more robust to the true interaction effect that is present in the data than the F test. For example, the power to detect an additive interaction effect is identical for the three different scenarios (see first column of Figure 1). This is because the Elston *et al.* test statistic for testing an additive interaction effect only uses the two extreme genotype groups (*aa* and *AA*), and so the position of the *Aa* group is irrelevant. The power to detect a dominant or recessive interaction does change as the underlying type of interaction effect changes. Nevertheless, as q changes the same trend in power is observed regardless of the true underlying interaction effect. For example, when n, σ^2 and μ_{ij} remain constant, it can be shown that the maximum power always occurs when q is 0.50, 0.39 and 0.61, when testing for an additive, dominant or recessive interaction effect, respectively. Figure 1 also shows that the Elston *et al.* test statistic always has low power when the allele frequency is extreme (close to 0 or 1), even if the correct type of interaction effect is tested for. This is easily explained by examining the form of these test statistics (4). As the number of observation in any group becomes small, the denominator of equation (4), in which the within-group variances are given equal weight, becomes large and power falls accordingly.

Table VI. Illustration of how the expected response probabilities, π_{ij}, may have (a) no interaction on the logit scale but has an interaction effect on the linear scale and (b) no interaction on the linear (probability difference) scale but has an interaction effect on the logit scale.

Linear model (Probability difference)	Logistic model
(a) *No interaction under the logistic model* ($\pi_{00}=0.05$, $\pi_{01}=0.13$, $\pi_{02}=0.29$, $\pi_{10}=0.29$, $\pi_{11}=0.52$, *and* $\pi_{12}=0.75$)	
$\pi_k = 0.520 + 0.231G_{1k} + 0.002G_{24} + 0.364T_k$ $+0.112G_{1k}T_k + 0.015G_{2k}T_k$	$\pi_k = 1 - \dfrac{1}{1+\exp\{0.1 + 1G_{1k} + 2T_k\}}$
(b) *No interaction under the linear model* ($\pi_{00}=0.05$, $\pi_{01}=0.35$, $\pi_{02}=0.65$, $\pi_{10}=0.30$, $\pi_{11}=0.60$, *and* $\pi_{12}=0.90$)	
$\pi_k = 0.6 + 0.3G_{1k} + 0.25T_k$	$\pi_k = 1 - \dfrac{1}{1+\exp\left\{\begin{matrix}0.585 - 1.522G_{1k} - 0.090G_{2k} + 1.567T_k\\ -0.259G_{1k}T_k - 0.271G_{2k}T_k\end{matrix}\right\}}$

Parameterization of the covariates, G_{1k}, G_{2k} and T_k are those of the general model as specified in Section 4.

By contrast, as shown in Figure 1, the F test corresponding to testing a specific type of interaction effect can be very sensitive to the type of interaction that is present in the data. However, the F test always has greater than or equal power compared to the Elston *et al.* statistic, if the correct model is used and the allele frequency is extreme. Moreover, the sensitivity of the F test to the true model can be overcome by fitting a general genetic model (see last column of Figure 1). This will work well when the number of genetic groups is small, but may be problematic when the number of genetic groups is large because of the rapid increase in degrees of freedom for the interaction terms, which would lead to a substantial decrease in power.

5.2. *Binary responses*

The Elston *et al.* and the GLM approaches are fundamentally different for the binary case because they use different scales of measurement. Interaction effects are scale dependent. For example, a model with gene and treatment main effects but no interaction on the logit scale, as used in the GLM approach, may generate a model with interaction on the difference in probability scale, as used by Elston *et al.*, and vice versa. Both cases are illustrated in Table VI where (a) there is a main treatment and gene effect but no interaction on the logit scale and (b) there is a main treatment effect and gene effect but no interaction on the probability difference scale. Figure 2 displays the power to detect an additive interaction effect for the Elston *et al.* and GLM approaches for the two scenarios displayed in Table VI. The plots show that the power to detect an interaction effect can be extremely different on the two scales. In other respects, the Elston *et al.* and LRT statistic for binary data display very similar patterns of power to their corresponding test statistics for the normally distributed data. For example, the power to detect an interaction effect for binary data using the Elston *et al.* statistic is always low if the allele frequency is extreme.

The power for the Elston *et al.* statistic (1) in Figure 2 has been calculated using variance estimate (a). Generally, there is relatively little difference between the power of tests using

Copyright © 2005 John Wiley & Sons, Ltd.

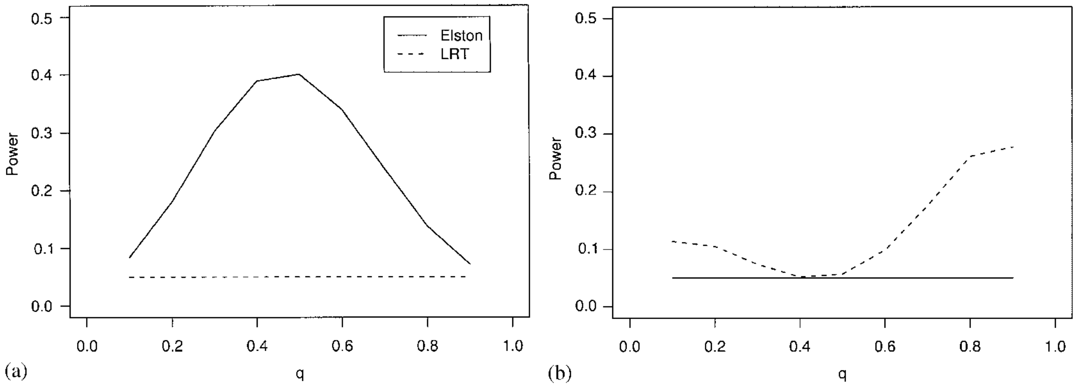

Figure 2. Power to detect an *additive* interaction effect for binary data using the Elston *et al.* and LRT statistics versus the allele frequency (q), for $n = 300$, $\alpha = 0.05$ and (a) $\pi_{00} = 0.05$, $\pi_{01} = 0.13$, $\pi_{02} = 0.29$, $\pi_{10} = 0.29$, $\pi_{11} = 0.52$, and $\pi_{12} = 0.75$ (no interaction on the logit scale), and (b) $\pi_{00} = 0.05$, $\pi_{01} = 0.35$, $\pi_{02} = 0.65$, $\pi_{10} = 0.30$, $\pi_{11} = 0.60$, and $\pi_{12} = 0.90$ (no interaction on the probability difference scale). HWE is assumed.

variance estimates (a) and (b) for test statistic (1) (see Figure 1 in Elston *et al.*), and hence the results using variance estimate (b) are not shown here.

5.3. Simulation study

In order to determine when large sample theory is adequate for the above approaches, the analytic power calculations for interaction effects were compared with those based on simulations for several scenarios—for different responses, allele frequencies and sample size. These were investigated at significance levels of 0.01 and 0.05, and assuming HWE. For binary data, the large sample theory appears to be reasonably accurate when the expected number of subjects within each genotype×treatment×response sub-group is at least one for the LRT approach and at least 10 for the Elston *et al.* approach. For normally distributed data, the *F* test and Elston *et al.* approaches appear to be adequate when the smallest expected number of subjects within a genotype×treatment sub-group is at least one. When these criteria are not met, or if the significance level is smaller than 0.01, the asymptotic theory may no longer adequately hold. It is therefore recommended that simulations be used in those cases to confirm the numerical calculations. The power calculations for all the examples presented in this paper were checked with simulations and were found to be similar to those based on the large sample theory.

6. DISCUSSION

In this paper we have presented two approaches for the analysis of pharmacogenetic trials, together with corresponding approaches to power calculation. We can see from the results presented above that both approaches have advantages and disadvantages. If the correct model is fitted and the allele frequency is extreme, so that at least one of the genotype groups is rare, then the GLM approach has at least equal and often much higher power compared to

Copyright © 2005 John Wiley & Sons, Ltd.

the Elston *et al.* approach. On the other hand, the Elston *et al.* approach is more robust to the mode of inheritance; for example, power to detect an interaction effect when fitting an additive model is the same whether the true model is additive, dominant or recessive. This is not true of the GLM-based approach, where the power of the GLM approach is very sensitive to misspecification of the model. We can overcome this problem for the GLM approach by fitting a general genetic model, thus avoiding any assumption about mode of inheritance. This is recommended when the number of genetic groups is small, but may not be feasible where the number of genetic groups is large because of the rapid increase in degrees of freedom for the interaction, which would lead to a substantial decrease in power. Fitting an additive genetic model, which works well under a range of true models, may be the best option unless we have very strong prior beliefs about the true genetic model.

A slightly different approach for the GLM method that has been presented in this paper is to always fit the full model, even when testing for a specific interaction effect. This is analogous to the approach employed by Tukey for dose-finding [7]. This should increase the power to detect an interaction effect if the wrong type of effect is tested for, since the variance estimate is decreased. It will, however, decrease the power slightly if the correct effect is tested for, due to the loss of two residual degrees of freedom.

The GLM approach has additional advantages in terms of flexibility. The power calculations can be easily extended to study designs with any number of genotype and treatment groups. The GLM approach can also test a number of other hypotheses of interest. For example, we can calculate the power to detect an interaction or treatment effect, adjusted for the main genetic effect, which reflects the gain/loss in power to detect a treatment effect by using genetic information. No such generalization of the Elston *et al.* approach is possible, this being limited to testing a gene or interaction effect of a single allele for two treatment groups or a dose–response relationship.

The interaction effect due to the scale of measurement is crucial for binary data. The results in this paper illustrate that the power to detect an interaction inevitably depends on the scale used. Even for models in which interactions are present on both scales, the relative magnitude of this interaction, and therefore the power to detect it, will vary with the measurement scale. It is therefore extremely important to consider on what scale the responses should be measured.

If asymptotic theory does not hold the simulation approach of calculating power used by Elston *et al.* is easy to apply to any test statistic of interest. However, the simulations under the null hypothesis for testing an interaction as proposed by Elston *et al.* are incorrect if there is a main allele effect, and it is not obvious how to correct this in their framework. This problem is easily solved in the GLM framework by simulating with a probability of success for each sub-group that has a main gene and treatment effect but no interaction effect.

Though we are not aware of any other published approaches to sample size calculations for a gene–treatment interaction; there are several methods available for the closely related problem of sample size calculations for gene–environment interactions. These are designed for genetic epidemiology studies where the outcome of interest is disease or no disease, and so assume a binary response. The most common approach has been to use a normal approximation to the odds ratio [8–10]. Two papers use methods similar to our GLM approach but these are in the context of case-control studies [11, 12].

When designing a trial to detect a gene–treatment interaction, there are a variety of approaches that can be taken with respect to choosing suitable input values for the sample size calculation(s). The approach taken depends somewhat on the prior knowledge available. For example, one may already know the expected response for each treatment group. Given these values, one could then examine the power to detect an interaction effect for a particular set of subgroup responses at a given allele frequency, or for a range of plausible scenarios with different subgroup frequencies and responses. As commented by a referee, it may also be useful to consider conducting a cross-over trial before designing a pharmacogenetic trial, in order to establish the magnitude of the subject-by-treatment interaction, as this provides an upper limit estimate of the gene-by-treatment interaction [13, 14].

ACKNOWLEDGEMENTS

This work was funded by Variagenics Inc., Cambridge MA, USA. The authors wish to thank Vincent Stanton Jr and David Clayton for their helpful discussions, and the referees for their constructive comments.

REFERENCES

1. Cardon LR, Idury RM, Harris TJ, Witte JS, Elston RC. Testing drug response in the presence of genetic information: sampling issues for clinical trials. *Pharmacogenetics* 2000; **10**:503–510.
2. Cuzick J. Interaction, subgroup analysis and sample size. In *Metabolic Polymorphisms and Susceptibility to Cancer*, Vineis P, Malatas N, Lang M, d'Errico A, Caporaso N, Cuzick J *et al.* (eds). Oxford University Press: Lyon, 1999.
3. Elston RC, Idury RM, Cardon LR, Lichter JB. The study of candidate genes in drug trials: sample size considerations. *Statistics in Medicine* 1999; **18**:741–751.
4. Hogg RV, Tanis EA. *Probability and Statistical Inference* (3rd edn). Macmillan Publishing Company: New York, 1983.
5. Dobson AJ. *An Introduction to Generalized Linear Models*. Chapman & Hall: London, 1990.
6. Johnson NL, Kotz S, Balakrishnan N. *Continuous Univariate Distributions*. (2nd edn). Wiley: Chichester, 1994.
7. Tukey JW, Ciminera JL, Heyse JF. Testing the statistical certainty of a response to increasing doses of a drug. *Biometrics* 1985; **41**:295–301.
8. Foppa I, Spiegelman D. Power and sample size calculations for case-control studies of gene–environment interactions with a polytomous exposure variable. *American Journal of Epidemiology* 1997; **146**:596–604.
9. Garcia-Closas M, Lubin JH. Power and sample size calculations in case-control studies of gene–environment interactions: comments on different approaches. *American Journal of Epidemiology* 1999; **149**:689–692.
10. Hwang S-J, Beaty TH, Liang K-Y, Coresh J, Khoury MJ. Minimum sample size estimation to detect gene–environment interaction in case-control designs. *American Journal of Epidemiology* 1994; **140**:1029–1037.
11. Longmate JA. Complexity and power in case-control association studies. *American Journal of Human Genetics* 2001; **68**:1229–1237.
12. Gauderman WJ. Sample size requirements for matched case-control studies of gene–environment interaction. *Statistics in Medicine* 2002; **21**:35–50.
13. Senn SJ. Individual therapy: new dawn or false dawn. *Drug Information Journal* 2001; **35**:1479–1494.
14. Kalow W, Tang BK, Endrenyi L. Hypothesis: comparisons of inter- and intra-individual variations can substitute for twin studies in drug research. *Pharmacogenetics* 1998; **8**:283–289.

Copyright © 2005 John Wiley & Sons, Ltd.