STATISTICS IN MEDICINE
Statist. Med. **18**, 741–751 (1999)

THE STUDY OF CANDIDATE GENES IN DRUG TRIALS: SAMPLE SIZE CONSIDERATIONS

ROBERT C. ELSTON[1]*, RAMANA M. IDURY[2], LON R. CARDON[2] AND JAY B. LICHTER[2†]

[1] *Department of Biostatistics and Epidemiology, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University, 2500 MetroHealth Drive, Cleveland, OH 44109, U.S.A.*
[2] *AxyS Pharmaceuticals Inc., 11099 North Torrey Pines Road, Suite 160, La Jolla, CA 92037, U.S.A.*

SUMMARY

With discovery of an increasing number of candidate genes that may affect inter-individual variability in response to drugs, the design of drug trials that incorporate their study has become relevant. We discuss the determination of sample size for such studies when the number of tests to perform is given, or, alternatively, the number of tests to perform when the sample size is given. In many cases, a uniformly most powerful test does not exist and normal approximations are not sufficiently accurate to determine sample size. We discuss briefly various tests of interest and we give simple examples to illustrate some of the problems that arise.

1. INTRODUCTION

In a typical drug trial, one randomizes subjects into dose groups and administers to each subject within a given group the same dose of the drug studied. After a predetermined amount of time, one measures an endpoint response and compares the mean response among groups. We take the variability of response within groups as random and we compare the mean group differences to this random inter-individual variability to evaluate their statistical significance. However, it is now recognized that among major factors determining inter-individual variability in response are variability in the levels of drug metabolizing enzymes (drug pharmacokinetics) or in receptor levels and activity (drug pharmacodynamics). Both of these factors, in turn, may have a major genetic component. There are approximately 100,000 genes in the human genome,[1] many of which may contribute to pharmacokinetic or pharmacodynamic variability. Consequently, there have been several large projects initiated to screen rapidly for polymorphic variants in medically relevant genes[2] and it is of interest to test whether the polymorphic variants of such candidate genes affect individual response to particular drugs.[3] With the rapidly increasing number of

* Correspondence to: Robert C. Elston, Department of Biostatistics and Epidemiology, Rammelkamp Center for Education and Research, MetroHealth Campus, Case Western Reserve University, 2500 MetroHealth Drive, Cleveland, OH 44109, U.S.A.
† Current address: Genset Corporation, 875 Prospect Street, Suite 206, La Jolla, CA 92037, U.S.A.

Contract/grant sponsor: National Institute of General Medical Sciences
Contract/grant number: GM 28356

Contract/grant sponsor: National Center for Research Resources
Contract/grant number: 1 P41 RR03655

CCC 0277–6715/99/060741–11$17.50

Received October 1997
Accepted May 1998

candidate genes discovered, the design of drug trials that incorporate their study has become increasingly relevant.[4] Furthermore, stratifying the dose groups on any such variants that do have an effect can increase the power of a study to detect dose differences.

In this paper we discuss designing a drug trial to test candidate genes, especially with respect to the sample size N required, when the response is a simple dichotomy – the subject either responds or does not. Many standard texts (for example, Schlesselman[5]) discuss sample size determination for studies involving a dichotomous response, but the study of candidate genes in drug trials warrants special consideration for several reasons. First, because there are at least three genotypes and two dose groups involved, the underlying distributions are often multinomial or product binomial rather than binomial, so that uniformly most powerful tests do not in general exist. Second, because we typically test more than one gene in any one trial, we have concern with the twin problems of determining the sample size N when there are n tests to perform, and how large n can be when the sample size N has already been specified. Third, the very small significance levels often needed to allow for the large number of genes tested can make the usual normal approximations for determination of sample sizes inappropriate. Fourth, in a similar vein, phase I and phase II clinical trials are often designed with relatively small sample sizes (often $N < 100$) which may be insufficient for assumptions of normality when stratified by genetic variants.

2. NOTATION AND TEST STATISTICS

The word 'gene' is used in genetics in two different senses: (i) the general kind of genetic material, DNA, that occurs at a particular place on chromosome and codes for a protein, or locus; (ii) the particular DNA composition that an individual has at that locus on one of two homologous chromosomes, or allele. The genotype an individual has at a locus is an unordered pair of alleles. Here our interest is in testing the effects of the various genotypes or alleles at one or more loci on drug response. In the traditional clinical study design, one administers to a fraction, k_j, of the sample drug dose j, $j = 0, 1, 2, \ldots, D-1$, the doses increasing with j and $j = 0$ usually corresponding to a placebo. The investigator determines the fractions k_j and they are often all equal. The quantities of primary interest in the study are the response rates at each dose, $j > 0$, which, when deemed significant larger than the response rate at $j = 0$, one takes as evidence for efficacious response.

Table I shows a simple extension of this design that incorporates differences in genotypes for a sample of N subjects. Consider an allele, A, at a particular locus that has relative frequency q in the population. Assuming there is random mating, negligible mutation and no direct selection with respect to the genotypes at this locus, a proportion, $c_0 = (1-q)^2$, of the population has 0 copies of the allele, a proportion, $c_1 = 2q(1-q)$, has one copy, and a proportion, $c_2 = q^2$, has two copies. In large samples, the expected number of individuals who have i copies of the allele and receive drug dose j is $N_{ij} = c_i k_j N$. Let R_{ij} denote the observed number of subjects in the ijth cell who respond to the drug, and π_{ij} the probability that a subject in that cell responds; then the expected number of responders to the drug in the ijth cell is $E(R_{ij}) = c_i k_j \pi_{ij} N$, and an estimate of the response rate π_{ij} is $p_{ij} = R_{ij}/N_{ij}$.

Any null hypothesis H_0 we wish to test is of the form

$$\sum_{i=0}^{2} \sum_{j=0}^{D-1} w_{ij} \pi_{ij} = 0$$

Table I. Numbers of subjects receiving drug dose j who have i copies of allele A; * indicates any other allele

Genotype	Genotype frequency	Dose 0	1	⋯	$D-1$	Total
**	$c_0 = (1-q)^2$	N_{00}	N_{01}	⋯	$N_{0,D-1}$	$c_0 N$
A*	$c_1 = 2q(1-q)$	N_{10}	N_{11}	⋯	$N_{1,D-1}$	$c_1 N$
AA	$c_2 = q^2$	N_{20}	N_{21}	⋯	$N_{2,D-1}$	$c_2 N$
Total		$k_0 N$	$k_1 N$	⋯	$k_2 N$	N

for a given set of coefficients w_{ij} such that $\sum_{i=0}^{2} \sum_{j=0}^{D-1} w_{ij} = 0$. We consider only one-sided tests, assuming that we choose the w_{ij} so that under any alternative hypothesis $\sum_{i=0}^{2} \sum_{j=0}^{D-1} w_{ij}\, \pi_{ij} < 0$. (We can accommodate a two-sided test by performing a second one-sided test, changing the signs of all the w_{ij}; we discuss the issue of multiple tests in the next section.)

We derive the required sample size N based on a statistic of the form:

$$t = \frac{\sum_{i=0}^{2} \sum_{j=0}^{D-1} w_{ij}\, p_{ij}}{\left\{ \sum_{i=0}^{2} \sum_{j=0}^{D-1} w_{ij}^2\, \hat{V}(p_{ij}) \right\}} \tag{1}$$

where $\hat{V}(p_{ij})$ is a consistent estimate of the variance of p_{ij} under the null hypothesis so that $t \sim \mathrm{N}(0, 1)$ in large samples. We consider two different variance estimates $\hat{V}(p_{ij})$ in (1):

(a) $p_{ij}(1 - p_{ij})/N_{ij}$;

(b) $p_j(1 - p_j)/N_{ij}$, where $p_j = \sum_i R_{ij} / \sum_j N_{ij}$.

The variance estimate (a) is consistent whatever the true values of the π_{ij}, while the estimate (b) is consistent under any hypothesis in which the π_{ij} do not depend on i, the number of allele copies a person has. The statistic that leads to the smaller value of N required indicates, for the particular alternative hypothesis of interest (the values of q, w_{ij} and π_{ij}), the more powerful of the two statistics to use. (A third variance estimate, $p_i(1 - p_i)/N_{ij}$, where $p_i = \sum_j R_{ij} / \sum_j N_{ij}$, is consistent under any hypothesis in which the π_{ij} do not depend on j, the drug dose to which an individual has been assigned; this estimate is of interest in testing dose effects, which we do not consider in this paper.)

3. SAMPLE SIZE DETERMINATION

In this section we consider sample size determination in general, deferring until later any consideration of the kinds of alternative hypotheses of possible interest. We first summarize sample size determination for the situation in which the sample size is so large that the assumption that t is normally distributed is a good approximation. In this situation $t \sim \mathrm{N}(0, 1)$ under the the null hypothesis and we reject the null hypothesis at the α-level when $t \leqslant z_\alpha$, where z_α is the usual α-fractile of the standard normal distribution. Under the alternative hypothesis, H_A,

$t \sim \mathrm{N}(E_A, 1)$, where

$$E_A = \frac{\sqrt{N} \sum_{i=0}^{2} \sum_{j=0}^{D-1} w_{ij}\pi_{ij}}{\sqrt{\left\{ \sum_{i=0}^{2} \sum_{j=0}^{D-1} w_{ij}^2 \frac{\pi_{ij}(1-\pi_{ij})}{c_i k_j} \right\}}} \tag{2}$$

and for power $1-\beta$ we must have

$$P(t - E_A \leqslant z_\alpha - E_A) \geqslant 1 - \beta$$

$$z_\alpha - E_A \geqslant z_{1-\beta}. \tag{3}$$

Substituting for E_{A} from (2) and solving for N, we arrive at the required sample size

$$N \geqslant \frac{(z_\alpha - z_{1-\beta})^2 \sum_{i=0}^{2} \sum_{j=0}^{D-1} w_{ij}^2 \frac{\pi_{ij}(1-\pi_{ij})}{c_i k_j}}{\sqrt{\left\{ \sum_{i=0}^{2} \sum_{j=0}^{D-1} w_{ij}\pi_{ij} \right\}}} \tag{4}$$

Because we will typically perform multiple tests, our taking the α-fractile of the standard normal distribution is unlikely to provide an adequate approximation for t_α, the α-fractile of t under H_0. Moreover, while our taking $z_{1-\beta}$ instead of the $(1-\beta)$-fractile of t under H_{A}, $t^*_{1-\beta}$, may often provide an adequate approximation because we are likely to use less extreme values of β, even this approximation may be inadequate when we use very small sample sizes – as in the case of phase I or phase II clinical trials. Therefore, we wish to calculate sample sizes using the exact quantities t_α and $t^*_{1-\beta}$, without relying on asymptotic theory. Here we describe an iterative simulation procedure for estimating these quantities.

Unlike the dosage condition of setting k_j experimentally, we cannot determine the frequencies c_i, $i = 0, 1, 2$, *a priori* because we do not know the individuals' genotypes until after the study is complete. Therefore, the number of individuals in each cell, N_{ij}, comes from a trinomial distribution with parameters $(k_j N, c_0, c_1, c_2)$. Furthermore, conditional on N_{ij} the number of responders to the drug, R_{ij}, comes from a binomial distribution with parameters (N_{ij}, π_{ij}). Thus, for any dose j, for fixed $k_j N$ we can simulate a sample of responders and non-responders for each of the three genotype classes under H_{A} by simulating from a six-nomial distribution with probabilities $c_0\pi_{0j}$, $c_0(1-\pi_{0j})$, $c_1\pi_{ij}$, $c_1(1-\pi_{ij})$, $c_2\pi_{2j}$ and $c_2(1-\pi_{2j})$. Under H_0, the proportion of responders does not depend on i, and so we simulate from a binomial distribution with parameters $k_j N$ and $\sum_{i=0}^{2} c_i\pi_{ij}$. We then obtain a sample of responders and non-responders in all categories by repeating this procedure, under H_0 or H_{A}, for each of the K doses. We explain this in more detail in the Appendix.

For any sample just described, we can calculate t as a function of the π_{ij} and the w_{ij} using (1), for either of the variance estimates, with the π_{ij} equal to their values under either H_0 or H_{A}. Thus, by simulating a large number of samples, we can estimate t_α and $t^*_{1-\beta}$. Now we reject the null hypothesis when $t \leqslant t_\alpha$, but, to ensure that we have power $1-\beta$, we also want $t \geqslant t^*_{1-\beta}$. In other words, we require that $t^*_{1-\beta} \leqslant t_\alpha$. We take the required sample size, N, as the smallest value that satisfies this condition. The Appendix provides details of a simple approximate iterative algorithm to achieve this.

In pharmacogenetic applications, not only will we perform multiple tests for a particular candidate locus, as discussed later, but usually we shall also test many candidate loci in the same trial. If we are to perform a total of n tests, then a simple procedure is to design the study on the assumption that we shall apply a Bonferroni correction, that is, take $\alpha = \alpha^*/n$, where α^* is the desired significance level after allowing for multiple independent tests. Because α is typically orders of magnitude smaller than α^* or β, we need to determine t_α accurately as the α-fractile of the distribution of t under the null hypothesis. In theory, it is possible to obtain the exact distributions of t and t^*, and thus the exact fractile values, by enumerating all possible outcomes. However, doing so in practice is combinatorially prohibitive even for sample sizes as small as $N = 50$, so we resort to using simulation as indicated above. We can determine the number of replicate samples required in the simulation by noting that if $\hat{\alpha}$ is an estimate of a binomial probability based on M trials, the width, W, of a 95 per cent confidence interval for α is approximately $4\sqrt{\{\alpha(1-\alpha)/M\}}$ so that $M = 16\alpha(1-\alpha)/W^2 \doteq 16\alpha/W^2$. Thus if we want $\alpha = 10^{-3}$ and we set $W = 5 \times 10^{-4}$, we should simulate about 6×10^4 replicate samples. In practice we find that taking $M = 50/\alpha$ or $50/(1-\beta)$, as appropriate, is usually adequate.

4. DETERMINING THE NUMBER OF TESTS

Sometimes the sample size for a drug trial has already been determined based on the mean group differences one wishes to detect, before one makes a decision to combine it with the study of candidate genes. In this situation it is of interest to determine how many candidate genes one can study and yet maintain a significance level equal to α^*, adjusted for multiple tests. We can do this using the same principles, but now N is known and so the procedure is simpler. We use simulation to obtain $t^*_{1-\beta}$ and the smallest α such that $t^*_{1-\beta} \leqslant t_\alpha$, as explained in the Appendix, and then set $n = \alpha^*/\alpha$ (rounded down to an integer).

5. EXAMPLES

One can program in some generality the method described above for determining N or n, allowing for input of sets of π_{ij} and w_{ij}, together with q, k_j ($j = 0, 1, \ldots, D-1$), α^*, β and either n or N. In this section we first discuss the various hypotheses of genetic interest. Then we examine a limited number of examples to illustrate that one or the other variance estimate can be better, and that the assumption of a normal distribution for t can lead to an erroneous estimate of the required sample size.

There are many alternative hypotheses of potential interest in clinical trials. The effect sizes that one wishes to direct, if such exist, are specified by the π_{ij} and depend on factors specific to the particular situation. On the other hand, the w_{ij} determine the kinds of alternatives one has interest in detecting. The simplest alternatives are those in which the allele in question has the same effect on response regardless of drug dose j. We test an additive allele effect by setting $w_{0j} = 1$, $w_{1j} = 0$, $w_{2j} = -1$, all j; a dominant allele effect by $w_{0j} = 2$, $w_{1j} = w_{2j} = -1$, all j; and a recessive allele effect by $w_{0j} = w_{1j} = 1$, $w_{2j} = -2$, all j. These three tests are correlated with each other and, because any two of these null hypotheses imply the third, they effectively count as two independent tests. Note that if there are only two alleles at a locus, say A_1 and A_2, a positive dominant effect of A_1 is the same as a negative recessive effect of A_2, and a positive dominant effect of A_2 is the same as a negative recessive effect of A_1. Here we use the term 'positive effect' to denote an increase in response, and the term 'negative effect' a decrease in response. Thus we

obtain a two-sided test for a dominant effect of allele A_1, for example, with two one-sided tests: one for a dominant effect of A_1 and one for a recessive effect of A_2. Furthermore, because there are only three genotypes, only two independent tests are possible.

When we believe that the candidate alleles being tested act by affecting response to the drug, so that π_{i0} (the response to placebo) does not depend on i (the number of candidate alleles that a subject has), these simple tests of overall additive, dominant or recessive effects are of lesser interest than are their interactions with drug dose. Let: a_i denote the values of w_{ij} given above to test for an additive allele effect; d_i those for a dominant allele effect; and r_i those for a recessive allele effect. Further, let $f_0, f_1, \ldots, f_{D-1}$ be the mean response rates for each dose (unweighted by genotype frequency), and let

$$e_j = Df_j - \sum_{j=0}^{D-1} f_j .$$

Then reasonable values of w_{ij} are $a_i e_j$ to detect an interaction of an additive allele effect with dose, $d_i e_j$ for an interaction of a dominant effect, and $r_i e_j$ for an interaction of a recessive effect. If the values of the f_j are completely unknown, then we could assume $f_j = j$ to determine sample sizes, though it is better to use the observed f_j when actually performing the tests. These three tests are correlated with each other, just as are the tests for each allele at a candidate locus, and they are again effectively two independent tests. One could decide *a priori* to perform the test for an additive allele interaction first, and then test for dominant and recessive interaction effects only to see if either of these offers a significantly better fit, in which case n, the number of tests to allow for, is smaller. Alternatively, for loci with large numbers of alleles, we could conservatively let n be twice the total number of alleles to test if we have interest only in interactions, or four times the number of alleles if we have interest in both allele effects and their interactions with dose. If all the loci are diallelic, n is half as large as that.

Figure 1 illustrates how the required sample size depends on the allele frequency q and the test performed. In each case we assume $D = 1$, $\alpha = 0{\cdot}05$, $1 - \beta = 0{\cdot}8$, and that under H_A there is the same response (0·1) for all genotypes in the placebo group, but the responses when the drug is administered are $\pi_{01} = 0{\cdot}1$, $\pi_{11} = 0{\cdot}5$ and $\pi_{21} = 0{\cdot}7$ (that is, an allele-by-dose interaction). Logarithms of the sample size required are plotted against the allele frequency when performing tests of additive, dominant or recessive effects, or interactions of such effects with dose, using each of the two variance estimates. In this particular case, because there are only two dose levels, we may use any two values of f_0 and $f_1 > f_0$, to determine the contrasts to test for interaction; although the w_{ij} differ, the value of t does not change. Also in this case, there is often little to choose between testing for an allele effect or for an allele-by-dose interaction.

Figure 2 shows two examples of how use of the normal approximation to determine sample size can lead to an incorrect estimate of the required sample size. We simulated the same model as for Figure 1, except that we now fixed q at 0·3 and we plotted log N against log α. When we use variance estimate (a) (the more powerful test for this situation), the normal approximation consistently underestimates the required sample size by about 20 per cent, whereas when we use variance estimate (b) it overestimates the required sample size. We have found empirically that with use of variance (a), the normal approximation usually underestimates N, but overestimates N when the allele frequency is very small or very large, that is, in those situations where the variances are more poorly estimated. The normal approximation can also overestimate or underestimate N with use of variance (b), but we observed no similar trend.

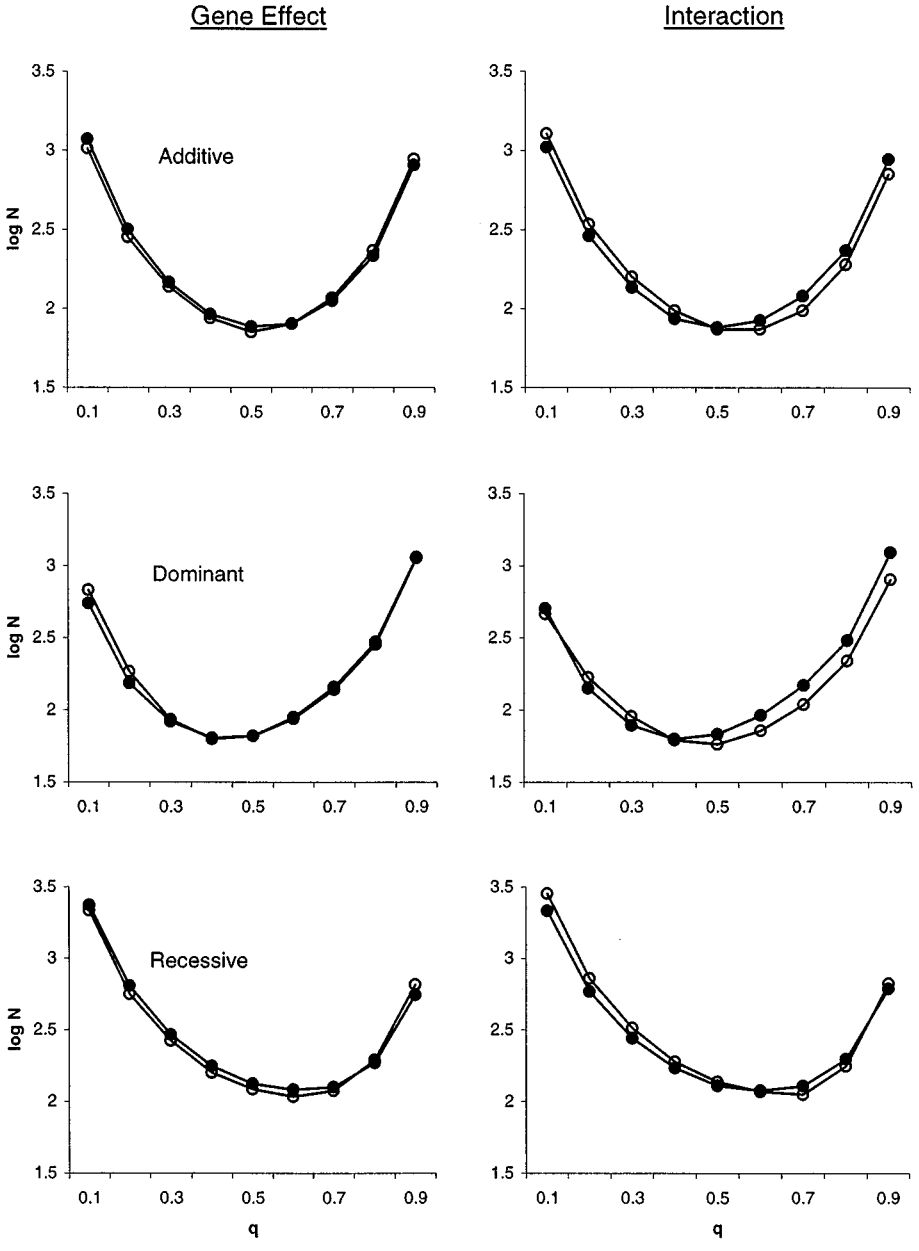

Figure 1. Comparisons of required sample size as a function of allele frequency for different tests with $\alpha = 0{\cdot}05$ and $1 - \beta = 0{\cdot}8$. All panels represent a two-dose design (one placebo, one dose) with response rates $\pi_{00} = \pi_{10} = \pi_{20} = \pi_{01} = 0{\cdot}1$, $\pi_{11} = 0{\cdot}5$ and $\pi_{21} = 0{\cdot}7$. Open and filled circles correspond to values obtained using variance (a) and (b), respectively. In descending order, the three left-side panels show log N required to test for additive allele effects ($w_{00} = w_{01} = 1$, $w_{10} = w_{11} = 0$, $w_{20} = w_{21} = -1$), dominant allele effects ($w_{00} = w_{01} = 2$, $w_{10} = w_{11} = -1$, $w_{20} = w_{21} = -1$), and recessive allele effects ($w_{00} = w_{01} = 1$, $w_{10} = w_{11} = 1$, $w_{20} = w_{21} = -2$), respectively. The three descending right-side panels show log N required to test for additive interaction ($w_{00} = w_{21} = -1$, $w_{10} = w_{11} = 0$, $w_{01} = w_{20} = 1$), dominant interaction ($w_{00} = -2$, $w_{01} = 2$, $w_{10} = w_{20} = 1$, $w_{11} = w_{21} = -1$), and recessive interaction ($w_{00} = w_{10} = -1$, $w_{01} = w_{11} = 1$, $w_{20} = 2$, $w_{21} = -2$), respectively. Logarithms are taken to the base 10

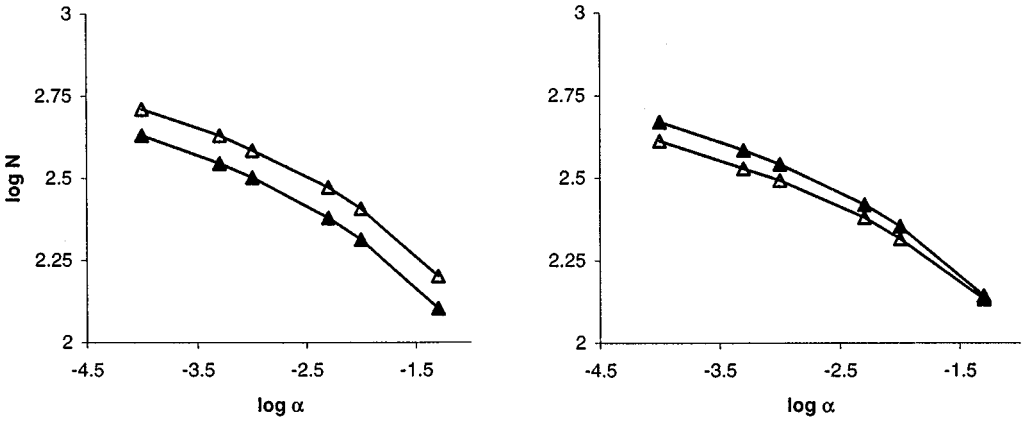

Figure 2. Comparison of sample size calculated using multinomial simulations (open triangles) versus normal approximations (filled triangles) for variance (a) (left panel) and variance (b) (right panel). In each panel, log N is plotted against log α to test for additive interaction reflecting one placebo and one dose group ($w_{00} = -1$, $w_{01} = 1$, $w_{10} = w_{11} = 0$, $w_{20} = 1$ and $w_{21} = -1$) when $\pi_{00} = \pi_{10} = \pi_{20} = \pi_{01} = 0{\cdot}1$, $\pi_{11} = 0{\cdot}5$ and $\pi_{21} = 0{\cdot}7$. Allele frequency, q, and power, $1 - \beta$, are fixed at 0·3 and 0·8, respectively. Logarithms are taken to the base 10

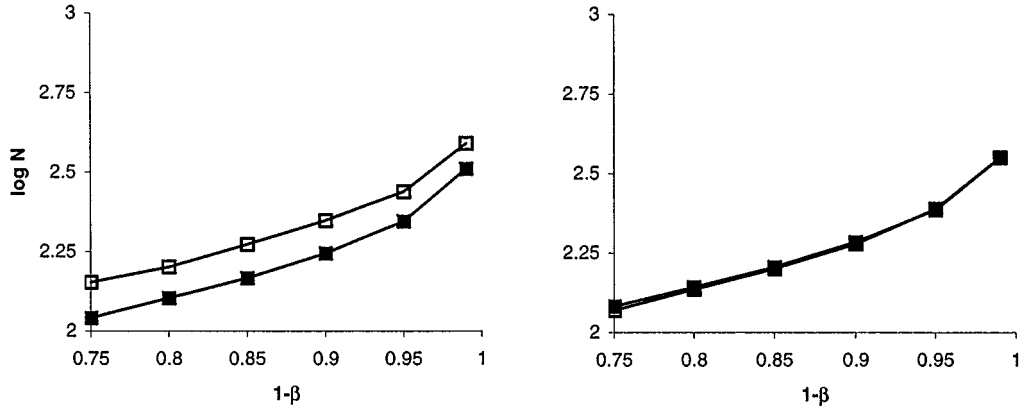

Figure 3. Plot of log N versus power to test for additive interaction in the same situation as for Figure 2. The left panel shows comparison of normal approximation (filled squares) with simulations (open squares) for variance (a); the right panel represents the same comparison for variance (b). α and q are fixed at 0·05 and 0·3, respectively

Figure 3 shows that, for the same example but fixing α at 0·05, the normal approximation seriously underestimates N for varying values of $1 - \beta$ with use of variance estimate (a), whereas it slightly overestimates N with use of variance estimate (b).

6. DISCUSSION

We have assumed that if we test multiple genes, whether multiple loci or multiple alleles at one or more loci, the same allele frequency q applies to each. In theory, we could allow for a different

frequency for each allele we wish to test and arrive at an overall sample size that allows exactly for such multiple testing. In practice this is complicated. Alternatively, we could derive the sample size assuming all the allele frequencies are equal to the smallest expected to occur in the sample, but this yields sample size requirements larger than really necessary. Also in theory, we could genotype a smaller fraction of each dose group for the more frequent alleles, and genotype all the members of each dose group only for the least frequent alleles. However, once we have collected a DNA sample from each subject, it is probably not worth the effort to subsample separately for each allele to be typed, especially if the genotyping is automated. Clearly, the allele frequency has a large effect when compared to the error incurred when using the usual asymptotic approximation to calculate sample size.

We have also assumed independence of the alleles at the different loci tested. To the extent that the loci tested are not independent, as in the case of different polymorphic variants at loci that are in 'linkage disequilibrium', the Bonferroni correction for multiple testing as described above may be too conservative and the values of N too large. In this situation we should decide *a priori* how many haplotypes (that is, combinations of alleles, one from each locus) we wish to test.

We note that we have discussed just one of many designs that incorporate the study of candidate genes into standard epidemiologic investigations. Recently, Umbach and Weinberg[6] noted how, in the case of incorporating candidate genes into a case-control study, greater power to detect gene–environment interaction is possible by using a log-linear model and by taking account of the fact that we expect genotypes and environmental factors to be distributed independently in the sample studied. Here we have studied a different design, one in which we assign cohorts of cases to different dose levels of drugs, but an analogous situation should hold. Provided we have interest in detecting interaction as a departure from a multiplicative model for allele and dose effects, rather than as a departure from an additive model, we can identify the non-responders with 'controls' and the responders with 'cases' and then perform the type of analysis indicated by Umbach and Weinberg[6] with the assumption that doses and genotypes are independent. Thus we would change the statistic (1) to one of the various test statistics they discuss, but we would otherwise determine sample size as indicated here. Also, analogous to their discussion of not typing 'controls' to detect allele–environment interaction, we could dispense with typing the placebo group to detect allele–drug-dose interaction – though we would not know if the effect detected is one of an overall allele effect or of an allele–drug interaction.

Finally, we note that the methods presented here are not limited to drug trials. For example (G. Weisner, personal communication), in the Bolton–Brush study conducted at Case Western Reserve University in the 1930s and 1940s, a large cohort of children had serial diagnostic radiography of the face and skeleton. There is a follow-up study under way to determine if the females have increased risk of breast cancer. Any such increased risk is likely slight, but may well be detectable among subgroups with specific alleles at candidate loci. In this situation the dose level is one of X-rays rather than drugs, and, although one could analyse the data as a dichotomy (presence or absence of breast cancer), an analysis of (censored) time to breast cancer would be more powerful. However, such a study shares two key components with the designs we have discussed: genotype classes are determined randomly rather than by design, and the multiple alleles that we want to test lead to a small desired value of α, and hence we need to allow for the fact that asymptotic approximation to determine sample size is probably inadequate. Furthermore, the large effect of allele frequencies would suggest it might be profitable to estimate them

from pilot studies before including the study of specific candidate genes in large-scale trials, and in such cases sample size requirements are more appropriately determined by the simulation approach we have described. These considerations are likely to become of general concern as the study of candidate genes becomes more widespread in epidemiologic and clinical investigations.

APPENDIX

In this Appendix, we first explain the details of a simulation based approximation scheme for estimating the x-fractile t_x of a distribution given x when N is known, or alternatively, estimating x given t_x, and then we give an iterative procedure to determine the sample size N.

Estimating t_x given x

Let N be the total number of people included in the drug trial. We simulate a single instance of a drug trial as follows. When $x = 1 - \beta$, divide the unit interval into six subintervals such that each interval is equal to the probability of one of the six outcomes of the six-nomial distribution. Let $N_j = k_j N$ be the predetermined number of individuals treated with dose j. For each dose j, generate N_j random numbers between 0 and 1. Depending on the subinterval in which each number falls, we place each of the N_j individuals into a genotype class as a responder or a non-responder. We can then compute $p_{ij} = R_{ij}/N_{ij}$ for each category, and we can do this for all doses to obtain a single trial set of outcomes p_{ij} under H_A, and hence a single instance of t. When $x = \alpha$, that is, we wish to simulate under H_0, the procedure is analogous except that we simulate from a binomial distribution for each dose j.

In each case we perform M simulated trials to obtain M values of t (in practice we set M to around $50/\alpha$ or $50/\beta$, as appropriate). We then sort the M values of t in ascending order to obtain the x-fractile t_x as the $\lceil xM \rceil$th entry of the sorted vector of values.

Estimating x given t_x

To solve the reverse problem of estimating x given t_x, we simulate trials as indicated above (in this case it is under H_0) and consider each trial a success if $t \leqslant t_x$, a failure otherwise. We simulate trials until we reach a certain number H of successful outcomes (in practice we set H around 50). Then, if M is the number of trials simulated so far, H/M gives an approximation of x. As a guard against simulating indefinitely when estimating very small values of x, we set an upper limit of M (say 10^8) and if we reach this limit, we conclude that x is too small to estimate within a reasonable amount of time.

Determining the sample size N

The following is a simple algorithm to find the desired sample size N starting with some small value N^0, such as 10. The values of N^i that follow are rounded so that all values of $k_j N^i$ are integers. We do the simulations indicated in the text with $N^1 = 2\ N^0$, to estimate t_α and $t^*_{1-\beta}$. If $t_\alpha < t^*_{1-\beta}$, we let $N^2 = 2\ N^1$ and repeat the process. We keep doing this until we reach a value N^m for which $t_\alpha \geqslant t^*_{1-\beta}$. At this point we know that the optimal value N_{opt} lies between $N_{\text{low}} = N^{m-1}$

and $N_{\text{high}} = N^m$. We therefore do a binary search as follows:

While $(N_{\text{low}} < N_{\text{high}})$ {
  $N_{\text{mid}} = (N_{\text{low}} + N_{\text{high}})/2$
  Simulate for $N = N_{\text{mid}}$
  If $(t_\alpha \geqslant t^*_{1-\beta})$ then {
    $N_{\text{high}} = N_{\text{mid}}$
  }
  Else {
    $N_{\text{low}} = N_{\text{mid}}$
  }
}
Stop, $N = N_{\text{mid}}$ is the optimal value.

ACKNOWLEDGEMENTS

This work was supported in part by U.S. Public Health Service research grant GM 28356 from the National Institute of General Medical Sciences, and resource grant 1 P41 RR03655 from the National Center for Research Resources.

REFERENCES

1. Fields, C., Adams, M. D., White, O. and Venter, J. C. 'How many genes in the human genome?', *Nature Genetics*, **7**,(3), 345–346 (1994).
2. Fodor, S.P.A. 'Massively parallel genomics', *Science*, **277**, 393–395 (1997).
3. Shpilberg, O., Dorman, J. S., Ferrell, R. E., Trucco, M., Shahar, A. and Kuller, L. H. 'The next stage: Molecular epidemiology', *Journal of Clinical Epidemiology*, **50**, 633–638 (1997).
4. Schork, N. J. and Weder, A.B. 'The use of genetic information in large-scale clinical trials: Applications to Alzheimer research', *Alzheimer Disease and Associated Disorders*, **10**, 22–26 (1996).
5. Schlesselman, J. J. *Case-Control Studies*, Oxford University Press, New York, 1982.
6. Umbach, D. M. and Weinberg, C. R. 'Designing and analyzing case-control studies to exploit independence of genotype and exposure', *Statistics in Medicine*, **16**, 1731–1743 (1997).

