

Designing Studies to Estimate the Penetrance of an Identified Autosomal Dominant Mutation: Cohort, Case-Control, and Genotyped-Proband Designs

Mitchell H. Gail,^{1*} David Pee,² Jacques Benichou,³ and Raymond Carroll⁴

¹*Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland*

²*Information Management Services, Rockville, Maryland*

³*Biostatistics Unit, University of Rouen Medical School, Rouen, France*

⁴*Department of Statistics, Texas A&M University, College Station*

One can obtain population-based estimates of the penetrance of a measurable mutation from cohort studies, from population-based case-control studies, and from genotyped-proband designs (GPD). In a GPD, we assume that representative individuals (proband) agree to be genotyped, and one then obtains information on the phenotypes of first-degree relatives. We also consider an extension of the GPD in which a relative is genotyped (GPDR design). In this paper, we give methods and tables for determining sample sizes needed to achieve desired precision for penetrance estimates from such studies. We emphasize dichotomous phenotypes, but methods for survival data are also given. In an example based on the BRCA1 gene and parameters given by Claus et al. [(1991) *Am J Hum Genet* 48:232–242], we find that similar large numbers of families need to be studied using the cohort, case-control, and GPD designs if the allele frequency is known, though the GPDR design requires fewer families, and, if one can study mainly probands with disease, the GPD design also requires fewer families. If the allele frequency is not known, somewhat larger sample sizes are required. Surprisingly, studies with mixtures of families of affected and non-affected

*Correspondence to: Mitchell H. Gail, Biostatistics Branch, National Cancer Institute, 6130 Executive Blvd., EPN/431, Rockville, MD 20892-7368. E-mail: gailm@epndce.nci.nih.gov

Received 17 October 1997; Revised 6 January 1998; Accepted 6 January 1998

This article is a U.S. government work, and, as such, is in the public domain of the United States of America.

Published 1999 Wiley-Liss, Inc.

probands can sometimes be more efficient than studies based exclusively on affected probands when the allele frequency is unknown. We discuss the feasibility and validity of these designs and point out that GPD and GPDR designs are more susceptible to a bias that results when the tendency for an individual to volunteer to be a proband or to be a subject in a cohort or case-control study depends on the phenotypes of his or her relatives. *Genet. Epidemiol.* 16:15–39, 1999. Published 1999 by Wiley-Liss, Inc.

Key words: bias; dichotomous phenotypes; family study; kin-cohort design; sample sizes for estimating penetrance; survival data

INTRODUCTION

Once a genetic mutation has been identified as affecting the risk of disease and the mutated gene has been isolated and techniques have been developed to detect it, several epidemiologic designs are available for estimating the risk of disease among carriers of the mutation. The dominant breast cancer genes BRCA1 and BRCA2 were identified by studying families with many affected women (“multiplex pedigrees”). Estimates of the lifetime risk (“penetrance”) of developing disease in mutation carriers from such families are high. For example, Easton et al. [1995] used linkage methods (rather than direct BRCA1 assays) to estimate that 85% of BRCA1 carriers would develop breast cancer by age 70. Their statistical methods took into account ascertainment of highly affected families by conditioning on the numbers of affected members in the pedigree. Nonetheless, some observers have questioned the applicability of these high penetrance estimates to the general population because these specifically selected pedigrees might have other genetic or environmental exposures that enhance the risk from BRCA1 and BRCA2 mutations. Indeed, a recent estimate of penetrance in a less selected population was only 56% [Struewing et al., 1997]. For this reason, there is a need for population-based studies to estimate the penetrance in the general population.

Struewing et al. [1997] recently reported on one such population-based design, which Wacholder et al. [1997], who developed the analytic methods used, called the kin-cohort design. Struewing et al. [1997] asked Ashkenazi Jews living in the Washington, D.C., area to volunteer for genotyping of the BRCA1 and BRCA2 genes. This population was selected because the prevalences of certain specific mutations were known to be elevated among Ashkenazi Jews. They then used information on the breast cancer history of relatives of these genotyped volunteers (the probands) to estimate the distributions of time to breast cancer among those with and without mutations. Their calculation of mutation-specific cumulative risk was based on the idea that the distribution of risk among relatives of a proband was a mixture over the unknown genotypes of the relatives of the gene-specific risk distributions, and the mixing probabilities for relatives could be calculated from the known genotype of the proband. Further details concerning the design and analytical approach for this study are given by Wacholder et al. [1997].

We use the term genotyped-proband design (GPD) in this paper, rather than kin-cohort design, to emphasize that the probands are genotyped and are selected at random, conditional on disease status. As in Struewing et al. [1997], we rely on the

family histories of first-degree relatives to provide information on penetrance, but, in addition, we allow information from genotyping the proband to contribute to the information on penetrance. Others such as Claus et al. [1991] and Whittemore et al. [1997] have used population-based samples of affected and unaffected probands and the family histories of their near relatives to estimate the penetrance of breast cancer, but their probands were not genotyped.

An ideal GPD would proceed by obtaining a random sample of N individuals with disease (case probands) and a random sample of M individuals without disease (control probands), genotyping the probands, and determining the cancer history (phenotype) of relatives. A variant on this design, GPDR, would be to genotype one or more relatives in addition to the proband. In this paper, we describe sample size calculations for GPD and GPDR needed to estimate the penetrance of a dichotomous trait with required precision, and we evaluate what ratios of N to M are most efficient. We outline similar calculations for estimating time-to-disease distributions with required precision.

The GPD and GPDR designs take advantage of the fact that the phenotypes of relatives of the probands can be obtained by reviewing their medical histories. Population-based case-control designs can also take advantage of retrospective evaluation of risk and genotype and yield estimates of gene-specific risk. We present an example that suggests that case-control designs may have comparable efficiency to the GPD design, and the case-control design can be more robust to certain selection biases than the GPD design. We also briefly consider the efficiency of a cohort design and its robustness to selection bias. A historical cohort study might be feasible, for example, if stored biological specimens permitted the genotyping of cohort members, some of whom may already have died at the time of the study.

We emphasize rare dominant alleles in our discussion, because several such genes have been identified as major cancer susceptibility genes, but the methods are easily extended to recessive or codominant autosomal genes.

The purpose of this paper is to define the required sample sizes and strengths and weaknesses of the cohort, case-control, GPD, and GPDR designs for estimating the penetrance of the mutant alleles with specified precision. We emphasize the binary case, though some results are also presented for time-to-disease data. Throughout we assume a simple genetic model in which the probability distribution of the phenotype is determined solely by the genotype. In particular, we assume that phenotypes within a pedigree are conditionally independent given corresponding genotypes. This commonly made assumption, while perhaps adequate for design considerations, would need to be examined critically when analyzing such data.

METHODS AND NOTATION

Binary Data

We suppose that each individual in the population has a genotype $g = 1$ or 0 according to whether the mutant allele is present or absent, as is appropriate for an autosomal dominant model. Assuming Mendelian genetics and Hardy-Weinberg equilibrium, we have $P(g = 1) = q^2 + 2q(1 - q)$ and $P(g = 0) = (1 - q)^2$, where q is the proportion of mutant alleles in the population. The binary outcomes, $Y = 1$ if dis-

eased and $Y = 0$ otherwise, define the penetrances $\phi_1 = P(Y = 1|g = 1)$ for carriers of the dominant mutation and $\phi_0 = P(Y = 1|g = 0)$ for non-carriers.

Cohort Design

To estimate ϕ_1 with required precision using the cohort design, we need to obtain a random sample of n_1 subjects with $g = 1$, where n_1 is large enough to make the width of the 95% confidence interval, 2Δ , acceptably small. In particular, we require

$$1.96 \{\phi_1(1 - \phi_1)/n_1\}^{1/2} = \Delta. \quad (2.1)$$

The problem is that in order to obtain n_1 subjects with $g = 1$, a very large number of members of the general population will need to be genotyped. Indeed, the expected number of required genotypes is $N = n_1/P(g = 1)$, which can be very large for rare mutations.

Case-Control Design

If $P(Y = 1)$ is known in the population, as in a population-based case-control study or in the study of a disease, such as breast cancer, for which population-based disease registries are available, it is possible to estimate ϕ_1 from Bayes theorem [Cornfield, 1951] as

$$\phi_1 = P(Y = 1)P(g = 1|Y = 1)\{P(Y = 1)P(g = 1|Y = 1) + P(Y = 0)P(g = 1|Y = 0)\}^{-1}. \quad (2.2)$$

The quantities $P(g = 1|Y = 1)$ and $P(g = 1|Y = 0)$ can be estimated by genotyping representative random samples of n cases and m controls, respectively, allowing computation of the estimate $\hat{\phi}_1$ from equation (2.2). Assuming $P(Y = 1)$ is known, the delta method [Rao, 1965, pp 319–325] yields the variance formula

$$\text{Var}(\hat{\phi}_1) = \{(1 - \varepsilon_1)/n\varepsilon_1 + (1 - \varepsilon_0)/m\varepsilon_0\} \{\phi_1(1 - \phi_1)\}^2, \quad (2.3)$$

where $\varepsilon_1 = P(g = 1|Y = 1)$ and $\varepsilon_0 = P(g = 1|Y = 0)$. Because ε_0 is usually much smaller than ε_1 , $\text{Var}(\hat{\phi}_1)$ can be minimized for a given total number of genotypes, $n + m$, by sampling more controls than cases. Indeed, the optimal allocation ratio is the square root of the odds ratio,

$$m/n = \{(\varepsilon_1/\varepsilon_0)(1 - \varepsilon_0)/(1 - \varepsilon_1)\}^{1/2}. \quad (2.4)$$

From equations (2.3) and (2.4), we can calculate the minimum number of genotypes $n + m$ needed to achieve required precision, namely a 95% confidence interval on ϕ_1 of width 2Δ . Analogous formulas hold for estimating ϕ_0 .

We note that case-control data can also be used to estimate ϕ_1 and ϕ_0 provided the allele frequency q is known, even if $P(Y = 1)$ is not known. This result follows from $P(Y = 1) = \pi_1\phi_1 + \pi_0\phi_0$, where $\pi_1 = P(g = 1) = 1 - \pi_0$. Indeed, solving the two equations $\phi_1 = P(Y = 1)\varepsilon_1/\pi_1$ and $(1 - \phi_1) = P(Y = 0)\varepsilon_0/\pi_1$ yields

$$\phi_1 = \varepsilon_1(\pi_1 - \varepsilon_0)/\pi_1(\varepsilon_1 - \varepsilon_0). \quad (2.5)$$

Genotyped Proband Design (GPD)

For design purposes we consider pedigrees consisting of a proband to be genotyped and up to two first-degree relatives whose medical histories will be ascertained. The data thus consist of Y_0 , g_0 , Y_1 , and Y_2 , where the subscript zero denotes the proband and subscripts 1 and 2 correspond to relatives. For example, consider a study of breast cancer in the pedigree in Figure 1. The proband is genotyped, and the breast cancer histories of her mother (Y_1) and sister (Y_2) are ascertained. Her father is not phenotyped because this disease has such low penetrance in men. Nonetheless, in computing the distributions of genotypes in the pedigree, conditional on g_0 , the possible genotypes of the father must be considered.

Assuming no inbreeding, non-assortative mating, Hardy-Weinberg equilibrium, and autosomal dominant inheritance, one can calculate the conditional distribution of genotypes $P(g_1, g_2 | g_0; q)$. We do this for pedigrees such as in Figure 1 by exhaustively enumerating the $3^4 = 81$ joint genotypes for the proband, two female relatives, and, in this case, the father. For example, letting AA, Aa, and aa denote homozygous

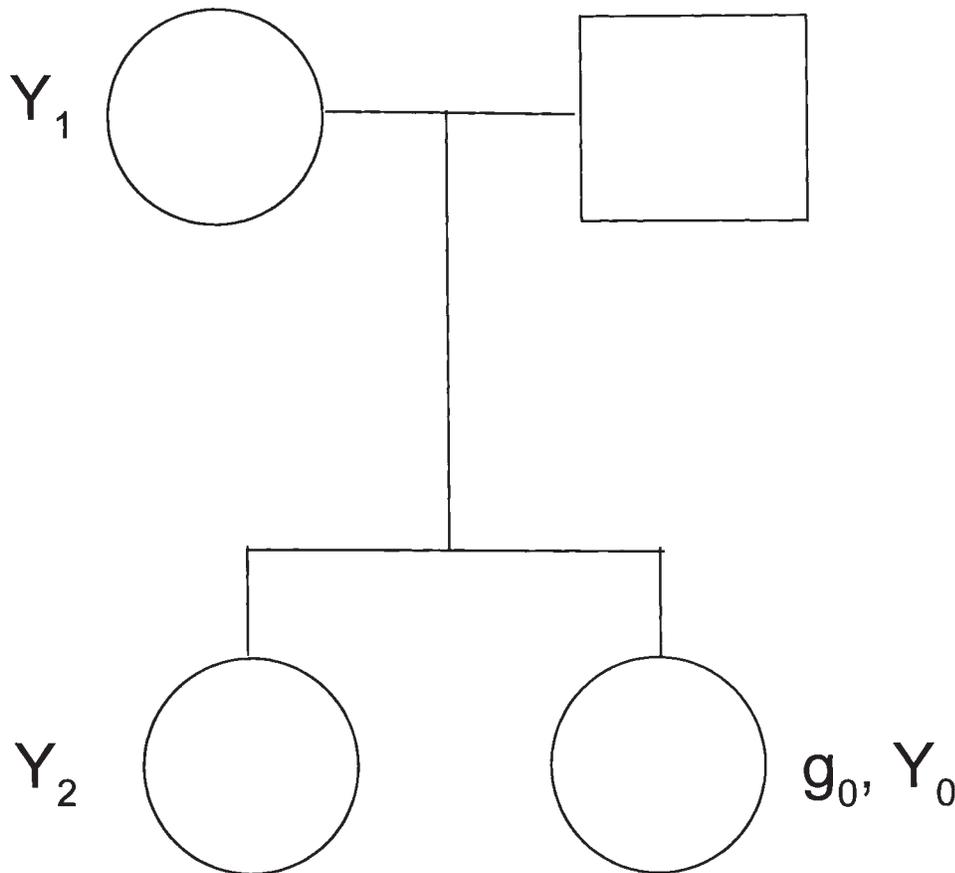


Fig. 1. Family of a female (circle) genotyped-proband with genotype g_0 and phenotype Y_0 . Phenotypic information on her mother (Y_1) and sister (Y_2) are also available.

carrier, heterozygote, and non-carrier, respectively, the probability that the father, mother, sister, and proband, respectively, have genotypes (aa), (Aa), (aa), and (Aa) is $(1 - q)^2 \times 2q(1 - q) \times (1/2) \times (1/2)$. Because events such as $g_0 = 1$ can be expressed in terms of events AA, Aa, and aa, the required conditional distribution $P(g_1, g_2 | g_0; q)$ can be calculated numerically as a function of q .

The likelihood contribution from a single family can be written as $L = L(Y_1, Y_2, g_0 | Y_0) = L_1 L_2$, where

$$L_1 = P(g_0 = 1 | Y_0) g_0 P(g_0 = 0 | Y_0)^{1-g_0} \quad (2.6)$$

is the contribution to the likelihood from genotyping the proband, and

$$L_2 = \sum_{g_1, g_2} \left\{ \prod_{i=1}^2 P(Y_i = 1 | g_i)^{Y_i} P(Y_i = 0 | g_i)^{1-Y_i} \right\} P(g_1, g_2 | g_0; q) \quad (2.7)$$

is the contribution from the phenotypes of the relatives. We are assuming in equation (2.7) that the responses of the relatives are conditionally independent given their genotypes. Note that no adjustments are needed for ascertainment bias, beyond what is implicit in this likelihood, because the ideal GPD is population-based, and probands are selected at random from among persons with ($Y_0 = 1$) or without disease ($Y_0 = 0$).

The quantities in equation (2.6) depend on the penetrances ϕ_1 , ϕ_0 , and on q through the relationships

$$\varepsilon_1 = P(g_0 = 1 | Y_0 = 1) = \pi_1 \phi_1 / (\pi_1 \phi_1 + \pi_0 \phi_0) \quad (2.8)$$

and

$$\varepsilon_0 = P(g_0 = 1 | Y_0 = 0) = \pi_1 (1 - \phi_1) / \{\pi_1 (1 - \phi_1) + \pi_0 (1 - \phi_0)\}. \quad (2.9)$$

The quantities $P(Y_i = 1 | g_i)$ in equation (2.7) are ϕ_1 or ϕ_0 according as $g_i = 1$ or 0.

The Fisher information matrix is obtained as the expectation of minus the cross derivatives of $\log(L)$ with respect to ϕ_1 and ϕ_0 , if q is assumed known, and with respect to ϕ_1 , ϕ_0 , and q otherwise (Appendix). Inverting the Fisher information matrix, I , yields I^{11} , the upper left-hand element of the inverse of I , which is asymptotically the variance of the maximum likelihood estimate of ϕ_1 . Hence the number of pedigrees required to achieve a confidence interval of ϕ_1 of width 2Δ can be computed. Details for writing the likelihood based on a cross-classification of families on g_0 , Y_1 , Y_2 , are given in the Appendix, where results for families with a single relative are also given. The results and methods in the Appendix are valid for any type of relative(s), such as two sisters of the proband, provided the appropriate conditional distributions $P(g_1, g_2 | g_0; q)$ are used.

If N case probands and M control probands are studied, we can represent the total information as

$$I = (N+M)\{\rho I(\text{case}) + (1 - \rho)I(\text{control})\}, \quad (2.10)$$

where $\rho = N/(N + M)$, I (case) is the information matrix from a single family with case proband and I (control) is the information matrix from a single family with control proband. To determine the most efficient fraction of families with case probands, we minimize the asymptotic variance of ϕ_1 , the appropriate element in the inverse of equation (2.10), with respect to ρ for $0 \leq \rho \leq 1$. For fixed ρ , the standard deviation, $SD(\phi_1)$, represents the precision per family studied in a mixed set of families with mixing proportion ρ . We use the notation $GPD(r,\rho)$ to represent a study of families with mixing proportion ρ in which each family yields phenotypes from r relatives of the proband.

The results in this paper will assume that the relative is the mother of the proband for $GPD(1,\rho)$ designs, and, for $GPD(2,\rho)$ designs, we assume that the relatives of the proband are the mother and sister. Other types of first-degree relatives yield almost identical results, however, because $P(g_1, g_2 | g_0)$ depends very little on the types of first-degree relatives with small allele frequencies.

Genotyped Proband Design With Supplemental Genotyping of Relatives (GPDR)

To explore the gains in information from genotyping relatives, we first consider the case of a family with a single relative in which both the proband and the relative are genotyped. The contribution to the likelihood from a single family of a case proband is L (case proband) =

$$\varepsilon_1^{g_0} (1 - \varepsilon_1)^{1-g_0} P(g_1 | g_0; q) \{ I(g_1 = 1) \phi_1^{Y_1} (1 - \phi_1)^{1-Y_1} + I(g_1 = 0) \phi_0^{Y_1} (1 - \phi_0)^{1-Y_1} \}, \quad (2.11)$$

where $I(\cdot)$ is an indicator function taking value 1 when the argument is true and zero otherwise. A similar expression is obtained for control proband families except ε_0 replaces ε_1 in equation (2.11). Equation (2.11) differs from the likelihood for the GPD design in two respects. First, there is a contribution, $P(g_1 | g_0; q)$, reflecting the genotyping of the relative. Second, the term in curly brackets in (2.11) is not a mixed distribution because only one of the two conditions $g_1 = 1$ and $g_0 = 0$ is true.

An alternative strategy is to genotype the relative only if the proband is a carrier. The likelihood for this strategy for a case proband is

$$\begin{aligned} L \text{ (case proband)} &= \varepsilon_1^{g_0} (1 - \varepsilon_1)^{1-g_0} \\ &\times [I(g_0 = 1) P(g_1 | g_0 = 1; q) \{ I(g_1 = 1) \phi_1^{Y_1} (1 - \phi_1)^{1-Y_1} + I(g_1 = 0) \phi_0^{Y_1} (1 - \phi_0)^{1-Y_1} \} \\ &+ I(g_0 = 0) P(Y_1 = 1 | g_0 = 0)^{Y_1} P(Y_1 = 0 | g_0 = 0)^{1-Y_1}]. \end{aligned} \quad (2.12)$$

The quantity $P(Y_1 = 1 | g_0 = 0) = \phi_1 P(g_1 = 1 | g_0 = 0) + \phi_0 P(g_1 = 0 | g_0 = 0)$.

For either strategy, the Fisher information is obtained by numerical differentiation after categorizing families according to g_0 , g_1 , and Y_1 (see Appendix).

As in equation (2.10), we study mixtures of families in which a proportion ρ have case probands, and we let $GPDR(\rho)$ denote a design with mixing proportion ρ . In this paper, we only analyze $GPDR$ designs with one relative, but we consider two strategies: always genotyping the relative or only genotyping the relative when $g_0 = 1$.

We assume that the relative in the GPDR design is the mother of the proband.

GPD Survival Data

Suppose the cumulative risk to age t of the disease of interest is

$$F_g(t) = 1 - S_g(t) = \phi_g [1 - \exp\{-(\lambda_g t)^{\alpha_g}\}] \quad (2.13)$$

for genotypes $g = 1$ (carriers) or $g = 0$ (non-carriers). The family (2.13) is quite flexible and corresponds to an improper Weibull distribution with lifetime risk (penetrance) ϕ_g as $t \rightarrow \infty$.

Consider a proband sampled at a calendar time C , and let a_0 and a_1 be the respective times from dates of birth to time C of the proband and a single relative, whom we assume to be the mother. Let d_0 and d_1 be the ages at death of the proband and relative. Note that $d_0 > a_0$ because the proband must be alive to be sampled, but $d_1 < a_1$ is possible. Let v_0 and v_1 be the ages when the disease of interest is incident in the proband and relative. Let $t_i = \min(a_i, d_i, v_i)$ and let $\delta_i = 1$ if $t_i = v_i$ and 0 if $t_i < v_i$ for $I = 0, 1$.

Conditional on g_0 , and assuming that other causes of death (and age) are independent of g_0 , the contribution to the likelihood from the relative is

$$\sum_{g_1} S_{g_1}(t_1) \{h_{g_1}(t_1)\}^{\delta_1} P(g_1 | g_0; q) G(t_1), \quad (2.14)$$

where the hazard is

$$h_g(t) = \phi_g \alpha_g \lambda_g^{\alpha_g} t^{\alpha_g - 1} \exp\{-(\lambda_g t)^{\alpha_g}\} / S_g(t), \quad (2.15)$$

and where $G(t)$ is the probability of surviving all non-breast cancer causes of death up to time t . Because G is assumed not to depend on g_1 , it does not affect estimation of penetrance through equation (2.14) and could be omitted.

Now consider a control proband with phenotype $Y_0 = (t_0 = a_0, \delta_0 = 0)$. To compute the contribution to the likelihood from genotyping the proband, namely

$$P(g_0 = 1 | Y_0)^{g_0} P(g_0 = 0 | Y_0)^{1-g_0}, \quad (2.16)$$

we note that

$$P(g_0, Y_0) = P(g_0) G(t_0) S_{g_0}(t_0). \quad (2.17)$$

Here $G(\cdot)$ is assumed to be independent of genotype, g_0 , and S_{g_0} can be estimated from cause-specific disease incidence rates in the presence of competing causes of death without the independence assumption [Prentice et al., 1978]. It follows that

$$P\{g_0 = 1 | Y_0 = (t_0, \delta_0 = 0)\} = P(g_0 = 1) S_1(t_0) \{P(g_0 = 1) S_1(t_0) + P(g_0 = 0) S_0(t_0)\}^{-1}. \quad (2.18)$$

Alternatively, a proband may have the disease of interest at an earlier age $t_0 = v_0$ but survive to age a_0 . In this case

$$P(g_0, Y_0) = P(g_0)G(t_0)S_{g_0}(t_0)h_{g_0}(t_0)J(a_0 - t_0; t_0), \quad (2.19)$$

where $J(u;v)$ is the probability that a person developing the disease of interest at age v will survive to age $v+u$. Note that both G and J are assumed to be independent of g_0 . The latter assumption may not always hold. From equation (2.19), we obtain

$$\begin{aligned} P\{g_0 = 1|Y_0 = (t_0, \delta_0 = 1)\} &= P(g_0 = 1)S_1(t_0)h_1(t_0)\{P(g_0 = 1)S_1(t_0)h_1(t_0) \\ &\quad + P(g_0 = 0)S_0(t_0)h_0(t_0)\}^{-1}. \end{aligned} \quad (2.20)$$

For each family, the likelihood is obtained by computing $P(g_0|Y_0)$ from (2.18) or (2.20), and then multiplying (2.16) by (2.14). We denote the logarithm of this product by $\ell(\phi_1, \phi_0, \lambda_1, \lambda_0, \alpha_1, \alpha_0; g_0, Y_0, Y_1)$. We can estimate the Fisher information, and hence the covariance of the parameter estimates, by obtaining a large random sample of data (g_0, Y_0, Y_1) , computing $\bar{\ell}$, the average value of ℓ in this sample, and obtaining the Hessian of $\bar{\ell}$ by numerical differentiation. We are interested in a particular function of ϕ_1 , λ_1 , and α_1 , namely the estimated cumulative risk to age 70, $F_1(70; \phi_1, \lambda_1, \alpha_1)$. The variance of this quantity is obtained by the delta method. Thus, we can estimate the number of families and genotypes needed to estimate $F_1(70)$ with required precision.

RESULTS

Example With Binary Outcome

We illustrate these calculations using parameters based on results by Claus et al. [1991], who estimated the lifetime probability of developing breast cancer among carriers of a hypothetical dominant gene as $\phi_1 = 0.92$. Estimates of the allele frequency $q = 0.0033$ and of the probability of developing disease in non-carriers, $\phi_0 = 0.10$, were also given. These calculations were based on studies of the phenotypes of members of families of cases and controls under age 55 from the Cancer and Steroid Hormone Study [Wingo et al., 1988]. Claus et al. [1991] did not measure gene status in cases and controls (BRCA1 and BRCA2 had not been cloned). Instead they performed a segregation analysis to investigate the plausibility of an autosomal dominant model, and they used the conditional distribution of the phenotypes of the relatives given the phenotypes of the probands (cases or controls) to estimate q , ϕ_0 , and ϕ_1 . In fact, they used survival data to estimate cumulative incidence curves, F_0 and F_1 , but, to illustrate our methods for binary data, we consider studies to estimate q , ϕ_0 , and ϕ_1 . We seek the number of families and genotypes needed to obtain a 95% confidence interval of width $2\Delta = 2 \times 0.05$ about the true penetrance $\phi_1 = 0.92$ using various designs (Table I).

In order to obtain required precision from a cohort study, $n_1 = 114$ carriers would need to be followed (see equation (2.1)). To obtain 114 carriers, one will need to genotype an expected $114/P(g = 1) = 114/\{.0033^2 + 2(.0033)(.9967)\} = 17,301$ women. The same number of women would need to be genotyped whether or not the allele frequency is known.

One often thinks of case-control studies as requiring many fewer subjects than cohort studies. As indicated in Table I, however, the case-control design requires

TABLE I. Numbers of Families and Genotypes Needed to Estimate the Penetrance of θ_1 With Precision $\pm 0.05^*$

Design	q unknown		q known	
	Families needed	Genotypes needed	Families needed	Genotypes needed
Cohort	N/A	17,301	N/A	17,301
Case-control	N/A	N/A	N/A	17,030 ^a
GPD				
1 Relative				
$\rho = 1$	26,851	26,851	14,439	14,439
$\rho = 0.5$	18,534	18,534	14,872	14,872
$\rho = 0.2$	16,119	16,119	15,658	15,658
$\rho = 0.1$	16,080	16,080	16,068	16,068
$\rho = 0$	4,534,930	4,534,930	16,584	16,584
2 Relative				
$\rho = 1$	13,418	13,418	8,808	8,808
$\rho = 0.5$	13,694	13,694	11,273	11,273
$\rho = 0.2$	14,234	14,234	13,778	13,778
$\rho = 0.1$	14,935	14,935	14,934	14,934
$\rho = 0$	61,028	61,028	16,338	16,338
GPDR (1 relative)				
Always genotype relative				
$\rho = 1$	3,549	7,098	3,231	6,462
$\rho = 0.5$	5,218	10,436	4,919	9,838
$\rho = 0.2$	7,367	14,734	7,289	14,578
$\rho = 0$	31,403	62,806	10,911	21,822
Genotype relative of $g_0 = 1$				
$\rho = 1$	3,940	4,167	3,907	4,132
$\rho = 0.5$	6,398	6,584	6,302	6,485
$\rho = 0.2$	12,347	12,495	9,970	10,090
$\rho = 0$	373,715	373,935	19,757	19,769

*The assumed parameters are $q = .0033$, $\theta_0 = .10$, and $\theta_1 = .92$. N/A, not applicable.

^aThis number is based on optimal sampling of 1,524 cases and 15,506 controls. The probability of disease, $P(Y=1)$, is assumed known from registry data. This is equivalent to knowing the allele frequency (see Methods and Notation).

genotyping 17,030 subjects, including 1,524 women with a personal history of breast cancer and 15,506 control women. Even larger numbers of genotypes would be required if the case/control ratio were nearer unity, because the ratio $1,524/15,506 = 1/10.17$ minimizes the required number of genotypes (see equation (2.4)). Large samples are required because the exposure (carrying a mutation) is so rare. The case-control design can only be used to estimate ϕ_1 if q or $P(Y = 1)$ is known.

The GPD design has very different sample size requirements and properties (Table I), depending on whether q is known or unknown, and depending on the proportion of probands who are cases (ρ). If q is known, required numbers of families and genotypes increase monotonically as ρ decreases, both for GPD(1, ρ) and GPD(2, ρ) designs. If all probands are non-diseased ($\rho = 0$), GPD(1,0) requires 16,584 families and GPD(2,0) requires 16,338 families, numbers that are comparable to requirements for the cohort and case-control designs. Using only case probands ($\rho = 1$) would reduce the needed sample sizes to 14,439 for GPD(1,1) and 8,808 for GPD(2,1).

If q is not known, much larger sample sizes are needed for GPD designs (Table I). Interestingly, $GPD(1,\rho)$ reaches a minimum required sample size 15,965 at $\rho = 0.14$ (data not shown), but the required numbers of families explodes to 4,534,930 as ρ tends to zero. The required sample sizes for $GPD(2,\rho)$ increase monotonically as ρ decreases and reach 61,028 at $\rho = 0$. Thus GPD designs can be more or less efficient than the cohort design if q is unknown, depending on the proportion of case probands.

The numbers of required families and genotypes can be reduced substantially if relatives can also be genotyped (Table I), especially for studies with a high proportion of case probands. With q known, only 3,231 families and 6,462 genotypes are needed for studies based on case probands only ($\rho = 1$) when all relatives are genotyped. Even further reductions in required genotypes can be achieved by genotyping the relative only when $g_0 = 1$, in which case 3,907 families and 4,132 genotypes are required. Similar efficiencies are seen even when q is unknown for GPDR designs (Table I), but these efficiencies diminish as the proportion of control probands increases, and vanish altogether when $\rho = 0$.

All these calculations are based on likelihoods that reflect information both from the genotype of the proband and from the phenotypes of the relatives. If we ignore the contribution to the likelihood, $P(g_0|Y_0)$ from genotyping the proband, the loss in precision for ϕ_1 can be considerable. For example, with q assumed known, the variance ratio comparing the full likelihood variance to the variance based only on the phenotype of the relative for the $GPD(1,1)$ design is 54%; if q is unknown, using only the phenotype of the relative is not sufficient to estimate ϕ_1 , ϕ_0 , and q . For the $GPD(1,0)$ design with q known, the variance ratio is 0.6%, and again, results are indeterminate if q is unknown and only the phenotype of the relatives is used. For the GPDR designs, losses from discarding information in $P(g_0|Y_0)$ are less severe. With q known and for case probands, the variance ratio for the GPDR with the relative always genotyped is 91%, and, with q unknown, over 99.9%. With control probands, these respective ratios are 35 and 99.8%. Thus, especially for GPD designs, a serious loss of information on ϕ_1 may result from failure to incorporate $P(g_0|Y_0)$ into the likelihood, and, if q is unknown, may result in indeterminate estimates.

To show that the nature of the first-degree relative makes very little difference in the GPD and GPDR designs, we recalculated some results in Table I with different types of first-degree relatives. The value 18,534 that arises in $GPD(1,0.5)$ with q unknown becomes 18,530 when the relative of the daughter proband is her sister, rather than her mother. Likewise, the value 13,694 for $GPD(2,0.5)$ becomes 13,702 when the two relatives are sisters instead of a sister and mother. If the relatives are genotyped as in GPDR designs, the required sample sizes are exactly as in Table I when the relative is a sister instead of a mother.

Table for Sample Size Calculations for GPD and GPDR Designs With q Unknown

To calculate the numbers of families needed to achieve a specified precision on ϕ_1 for a range of parameter values ϕ_0 , ϕ_1 , ρ , and q , we approximate the standard deviation (SD) obtained for ϕ_1 from data on a single family by a regression model. To achieve a desired standard deviation SD_0 , one calculates the required number of families as $(SD/SD_0)^2$. These tabulations pertain to the case in which q is assumed unknown.

The logarithm of the standard deviation of $\hat{\phi}_1$ decreases smoothly in $\log(q)$ for a range of values of q near $q = 0.01$ (Fig. 2). The data in Figure 2 correspond to GPD(1, ρ) designs. Note that a quadratic fits each locus well. The largest standard deviations are found for $\rho = 1$, in accord with Table I. Such loci can be accurately described by the equation

$$\log(\text{SD}) = \mu + \beta_1 \ln(100q) + \beta_2 \{\ln(100q)\}^2. \quad (3.1)$$

In equation (3.1), the parameters μ , β_1 , and β_2 depend on the design used and on ϕ_0 , ϕ_1 , and ρ . We used unweighted least squares to fit such a regression over the values $q = 0.001, 0.002, 0.003, 0.004, 0.005, 0.006, 0.007, 0.008, 0.009, 0.010, 0.025, 0.050, 0.075, 0.100$. If a locus such as in Figure 2 reaches a minimum in the range $0.001 < q < 0.100$, the value of q at the minimum is used as the upper limit of the range of q values to which the regression is fitted. We mention this upper limit in footnotes to Tables II–V, which contain the parameters μ , β_1 , and β_2 in equation (3.1).

For a given set of parameter estimates, $\hat{\mu}_1$, $\hat{\beta}_1$, and $\hat{\beta}_2$, the estimated SD is obtained by exponentiating the right-hand side of equation (3.1). Note that $\exp(\hat{\mu})$ corresponds to the estimated SD at allele frequency $q = 0.01$.

To illustrate the use of Table II, consider $\phi_0 = 0.10$, $\phi_1 = 0.95$, $\rho = .1$, and $q = 0.0033$, which are similar to the parameters in Table I. Table II pertains to the GPD(1, ρ) design in which phenotypic information is available from the mother of the proband. From equation (3.1), $\text{SD} = \exp[.41391 - .49778 \ln(.33) - .00006 \{\ln(.33)\}^2] =$

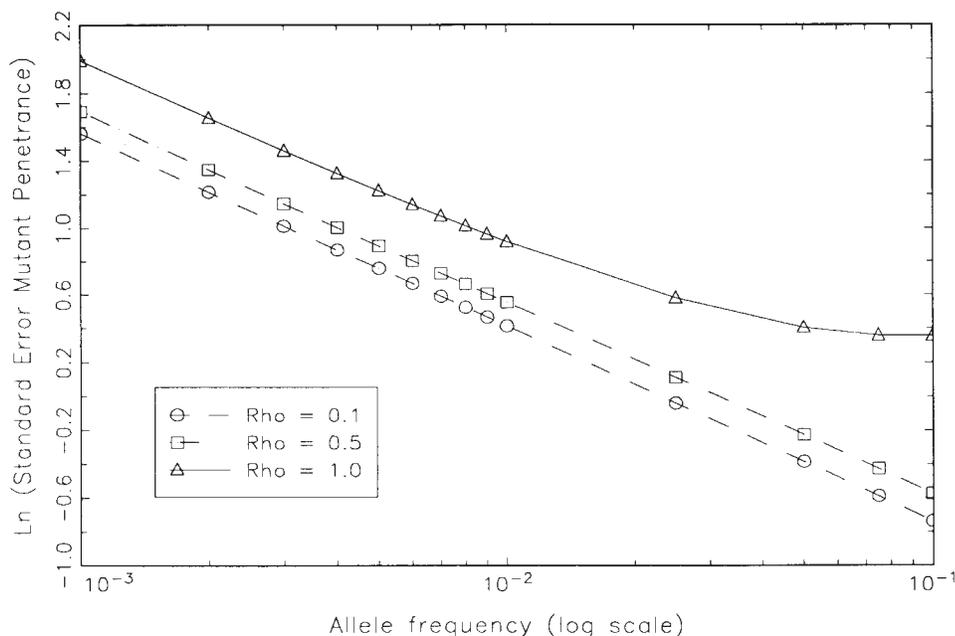


Fig. 2. Plot of the natural logarithm of the standard error of $\hat{\phi}_1$ against the allele frequency, q , on a log scale. Data are for the GPD(1, ρ) design with $\phi_1 = .95$, $\phi_0 = .10$, and $\rho(\text{Rho}) = 0.1, 0.5$ or 1 , as indicated. The standard error of $\hat{\phi}_1$ corresponds to the information from an average family drawn from the mixture of families. The allele frequency is assumed unknown.

TABLE II. Regression Coefficients to Approximate the Standard Deviation of $\hat{\theta}_1$ Based on One Family for the GPD (1, ρ) Design With Unknown Allele Frequency*

(θ_0, ρ)		θ_1			
		.25	.50	.75	.95
.01	1	.17365 - .29099 .05726 ^b	.23138 - .21940 .06341 ^a	.25361 - .15634 .07293 ^a	.24613 - .11257 .07988 ^a
	.5	.32539 - .34371 .03581	.45651 - .29970 .03544	.46699 - .28512 .02651	.28385 - .35307 .00213
	.1	.66758 - .35524 .01884	.90728 - .35977 .01546	.88377 - .38210 .00625	.38958 - .46399-.00555
.05	1	.86755 - .42358 .02797	.79380 - .36900 .04582	.72982 - .33437 .05215 ^c	.66449 - .31859 .05221
	.5	.64122 - .43976 .01953	.84035 - .43113 .02094	.81568 - .42465 .01912	.49019 - .46286 .00527
	.1	.83152 - .43796 .01807	1.05701 - .43296 .01773	1.00070 - .44812 .01286	.41591 - .49307-.00005
.10	1	1.24600 - .46342 .01428	1.11687 - .42470 .02803	1.01522 - .39416 .03813	.92571 - .37248 .04532
	.5	.74966 - .46377 .01293	.98937 - .46454 .01230	.96290 - .46493 .01060	.55504 - .48992 .00119
	.1	1.0019 - .47146 .00987	1.16943 - .45890 .01289	1.06397 - .46465 .01085	.41391 - .49778 .00006
.20	1	1.67030 - .48998 .00148	1.47361 - .46440 .01373	1.33162 - .44383 .02116	1.21367 - .42988 .02621
	.5	.86036 - .48030 .00727	1.09638 - .48221 .00654	1.05830 - .48762 .00431	.56524 - .50374-.00203
	.1	1.21008 - .49348 .00279	1.32188 - .48060 .00639	1.14647 - .47755 .00778	.39291 - .49955-.00006
.35	1		1.77464 - .48754 .00421	1.58754 - .47291 .01012	1.43301 - .46416 .01382
	.5		1.14668 - .48831 .00428	1.07085 - .49444 .00211	.50316 - .50787-.00317
	.1		1.45464 - .49378 .00182	1.23243 - .48691 .00475	.35055 - .49919 .00034

*The relative of the proband is the mother. The coefficients are μ , β_1 and β_2 , respectively, in equation (3.1).

^aUse only for $q \leq .025$; ^bUse only for $q \leq .05$; ^cUse only for $q \leq .075$. Otherwise valid for $q \leq 0.10$.

TABLE III. Regression Coefficients to Approximate the Standard Deviation of $\hat{\theta}_1$ Based on One Family for the GPD (2, ρ) Design With Unknown Allele Frequency, q^*

(θ_0, ρ)		θ_1			
		.25	.50	.75	.95
.01	1	-.17213 - .28847 .05821 ^b	-.11364 - .21564 .06479 ^a	-.08844 - .14710 .07626 ^a	.10998 - .14612 .06523 ^d
	.5	.03244 - .35027 .03617	.14646 - .29017 .03822	.16437 - .26245 .03277	.05100 - .30741 .01094
	.1	.44172 - .38403 .02068	.67138 - .36343 .01750	.68399 - .36505 .00851	.30727 - .44189 .00601
.05	1	.52098 - .42413 .02786	.44668 - .36829 .04630	.38321 - .34452 .04599 ^b	.31885 - .31377 .05435 ^b
	.5	.46537 - .45805 .01438	.59941 - .43039 .02104	.56785 - .41099 .02316	.33087 - .43957 .01101
	.1	.70389 - .46227 .01179	.90921 - .45432 .01359	.87477 - .45972 .01085	.37267 - .49356-.00028
.10	1	.89939 - .46474 .01376	.76969 - .42534 .02769	.66803 - .39377 .03849	.58159 - .37750 .04098 ^c
	.5	.62686 - .48050 .00734	.80023 - .46859 .01064	.76296 - .45840 .01246	.44311 - .47819 .00444
	.1	.88021 - .48532 .00528	1.04745 - .47782 .00725	.96956 - .48274 .00599	.38527 - .50355-.00169
.2	1	1.32098 - .49489 .00275	1.12620 - .46600 .01313	.98442 - .44500 .02070	.86696 - .42977 .02629
	.5	.77317 - .49265 .00287	.96021 - .48956 .00387	.91576 - .48808 .00388	.49766 - .50035-.00109
	.1	1.07205 - .49879 .00017	1.20355 - .49282 .00240	1.07388 - .49410 .00222	.37523 - .50711-.00270
.35	1		1.42684 - .48931 .00330	1.23956 - .47518 .00958	1.08590 - .46534 .01315
	.5		1.04237 - .49626 .00110	.97055 - .49897 .00032	.46272 - .50817 .00342
	.1		1.31559 - 50011-.00072	1.16321 - .49817 .00048	.33878 - .50565-.00214

*The relatives of the proband are the mother and sister. The coefficients are μ , β_1 and β_2 , respectively, in equation (3.1).

^aUse only for $q \leq .025$; ^bUse only for $q \leq .05$; ^cUse only for $q \leq .075$. Otherwise valid for $q \leq 0.10$.

TABLE IV. Regression Coefficients to Approximate the Standard Deviation of $\hat{\theta}_1$ Based on One Family for the GPDR Design With Unknown Allele Frequency, q^*

(θ_0, ρ)		θ_1			
		.25	.50	.75	.95
.01	1	.04218 - .31835 .04790	-.00919 - .25109 .05251	-.24355 - .21160 .05178	-.97441 - .18922 .05013
	.5	.24114 - .32484 .04228	.26471 - .27266 .04577	.05106 - .23673 .04463	-.67424 - .22344 .03924
	.1	.59442 - .33658 .02535	.75847 - .32250 .02792	.60074 - .30202 .03003	-.11811 - .30765 .02644
.05	1	.61590 - .43884 .02148	.50236 - .39550 .03362	.21258 - .36359 .04055	-.55526 - .34320 .04421
	.5	.51949 - .42359 .02544	.63147 - .40986 .02925	.42265 - .39046 .03217	-.31168 - .38041 .03090
	.1	.70094 - .42362 .02311	.88504 - .41313 .02554	.74983 - .41172 .02744	.05155 - .43526 .01905
.10	1	.86037 - .46835 .01219	.75962 - .43907 .02162	.46515 - .41529 .02846	-.30917 - .39916 .03266
	.5	.60726 - .44907 .01836	.77196 - .44593 .01949	.59474 - .43777 .02104	-.12581 - .43294 .02021
	.1	.80503 - .45787 .01500	.96292 - .44375 .01868	.81729 - .44161 .02039	.11247 - .46537 .01231
.20	1	.06993 - .48489 .00557	1.00441 - .46808 .01197	.72026 - .45280 .01723	-.05207 - .44164 .02078
	.5	.68836 - .46674 .01256	.87291 - .46611 .01278	.72844 - .46702 .01235	.02976 - .46924 .01035
	.1	.90868 - .47913 .00841	1.05165 - .46763 .01156	.88445 - .46235 .01389	.14315 - .48097 .00762
.35	1		1.17695 - .48244 .00663	.91090 - .47288 .01034	.14635 - .46542 .01297
	.5		.92100 - .47380 .00996	.78560 - .47824 .00856	.09913 - .48677 .00479
	.1		1.11048 - .48029 .00737	.93797 - .47462 .00964	.13745 - .48596 .00583

*The relative of the proband is the mother, who is also genotyped. The coefficients are μ , β_1 and β_2 , respectively, in equation (3.1). Valid for $q \leq .10$.

TABLE V. Regression Coefficients to Approximate the Standard Deviation of $\hat{\theta}_1$ Based on One Family for the GPDR Design With Unknown Allele Frequency, q , in Which the Relative Is Genotyped Only If the Proband Is a Carrier*

(θ_0, ρ)		θ_1			
		.25	.50	.75	.95
.01	1	-.52331 - .31874 .04765	-.19885 - .25135 .05241	-.27511 - .21178 .05171	-.97031 - .18936 .05008
	.5	-.20673 - .33100 .04300	.12682 - .26088 .04894	.05189 - .21889 .04965	-.64913 - .20196 .04706
	.1	.44563 - .38707 .02420	.85273 - .31018 .03145	.81524 - .24839 .03803	.10430 - .20567 .04703
.05	1	.13986 - .44118 .02067	.36349 - .39670 .03313	.21207 - .36442 .04022	-.53003 - .34386 .04395
	.5	.37517 - .45094 .01676	.63730 - .40626 .02924	.49465 - .37488 .03689	-.25092 - .36502 .03677
	.1	.78840 - .47807 .00732	1.23538 - .44155 .01631	1.17985 - .39816 .02783	.38241 - .36913 .03755
.10	1	.48549 - .47211 .01055	.68114 - .44132 .02074	.50237 - .41687 .02784	-.25860 - .40043 .03216
	.5	.62639 - .47953 .00830	.89488 - .45050 .01754	.73607 - .42726 .02416	-.02751 - .42214 .02437
	.1	.87360 - .49146 .00354	1.37166 - .47477 .00840	1.34421 - .44624 .01670	.51843 - .42415 .02511
.2	1	.87278 - .49085 .00313	1.03909 - .47201 .01047	.82846 - .45563 .01610	.04598 - .44398 .01986
	.5	.84520 - .49250 .00266	1.14585 - .47838 .00800	.97794 - .46410 .01298	.18590 - .46318 .01277
	.1	.92144 - .49466 .00236	1.45618 - .49049 .00335	1.47171 - .47817 .00753	.64570 - .46202 .01445
.35	1		1.36909 - .48805 .00437	1.11673 - .47721 .00863	.30890 - .46906 .01154
	.5		1.33423 - .49059 .00308	1.16827 - .48271 .00633	.32691 - .48305 .00620
	.1		1.48734 - .49420 .00180	1.53596 - .49209 .00250	.76203 - .48034 .00639

*The relative of the proband is the mother. The coefficients are μ , β_1 and β_2 , respectively, in equation (3.1).
Valid for $q \leq .10$.

$\exp(.96571) = 2.62664$. To obtain a desired standard deviation of $SD_0 = 0.05/1.96 = .02551$, we therefore require $(SD/SD_0)^2 = 10,602$ families. This compares with the number 16,080 in Table I for $\phi_1 = 0.92$, because SD decreases with increasing ϕ_1 in this range. To try to estimate the number of families required for $\phi_1 = .92$ from Table II, we calculated that 37,133 families were required for $\phi_1 = 0.75$, $\phi_0 = 0.10$, $\rho = 0.1$, and, using linear interpolation between $\phi_1 = 0.75$ and $\phi_1 = 0.95$, we obtained an estimate of 14,582 families needed for $\phi_1 = 0.92$. Thus, reasonable approximations may be obtained by linear interpolation from calculations based on Table II.

The calculations of SD derived from Table II are themselves quite accurate. For example, the $SD = 2.62664$ calculated above is quite close to the exact value, 2.62534.

We assessed the accuracy of these regressions in terms of the relative error, $100 \times |\hat{SD} - SD|/SD$, where \hat{SD} is estimated from the regression and SD is the true value. Over the valid ranges of q indicated in the footnote to Table II, the maximum relative error was 4.6%, but most relative errors were less than 1% for $0.002 \leq q \leq 0.01$. The maximum relative errors in Tables III–V were, respectively, 4.6, 2.6, and 2.9%, and most relative errors were much smaller.

Survival Data

To estimate the precision with which the cumulative incidence of disease can be estimated for carriers and non-carriers from the GPD design, we simulated data with similar cumulative incidence functions as those estimated by Claus et al. [1991]. In particular, we used Weibull models $F_1(t)$ and $F_0(t)$ given by equation (2.13) with $\phi_1 = .92$, $\lambda_1 = .016$, and $\alpha_1 = 4.047$ and with $\phi_0 = .10$, $\lambda_0 = .013$, and $\alpha_0 = 5.148$. The means and variances of F_1 and F_0 match those of the normal survival models in Table III of Claus et al. [1991]. In particular, F_1 has a mean of 55.4 years and standard deviation of 15.4 years, and F_0 has a mean of 69.0 years and standard deviation of 15.4 years. We used the allele frequency of $q = 0.0033$ from Claus et al. [1991]. We used a bivariate normal distribution of (potential) ages of the proband, a_0 , and mother, a_1 , modeled on data from the study by Struewing et al. [1997]. The means of a_0 and a_1 were 51.96 and 80.35 years, and the covariances were $\sigma_{00} = \sigma_{01} = 201.6$ and $\sigma_{11} = 381.77$. Non-breast cancer related death times, d_0 and d_1 , were generated from a piecewise constant hazard model (5-year intervals) of U.S. mortality rates, excluding breast cancer. Ages at cancer incidence v_0 and v_1 were generated from F_0 and F_1 . For probands who developed breast cancer before age a_0 , we generated an exponentially distributed survival time with mean 20 years. If d_0 and v_0 plus this survival time both exceeded a_0 , the data from the proband and her mother were used. Otherwise, we discarded this pair, because such a pair corresponds to a proband who had died before the study. The quantities Y_0 and Y_1 were then calculated as described in Methods and Notation from (a_0, d_0, v_0) and (a_1, d_1, v_1) .

One million simulations, each containing 5,000 families with a proband and mother, were used to estimate the information matrix, as described in Methods and Notation. These simulations were used to estimate the information matrix and not to study the statistical properties of \hat{F}_1 or \hat{F}_0 . Values of $F_1(t)$ at ages 40, 50, 60, and 70 were, respectively, .151, .330, .556, and .757, and, from the information matrix, we calculated respective standard errors of \hat{F}_1 from the GPD design with 5,000 families of .053, .073, .093, and .111. Corresponding values $F_0(t)$ were .004, .012, .027, and .050, with respective standard errors of \hat{F}_0 of .001, .001, .002, and .003. The logit

transform reduces skewness in \hat{F}_1 and \hat{F}_0 when F_1 and F_0 are near 0 or 1. Using the delta method for the logit transform $\log \{F_1/(1 - F_1)\}$, we obtained an expected confidence interval on $F_1(70)$ of (.487, .911). Thus, even estimates based on 5,000 families are subject to considerable random uncertainty, as was noted by Struewing et al. [1997].

Unreported data confirm that the simulation approach to estimating the Fisher information matrix yields covariance estimates for underlying parameters that are in good agreement with empirical estimates of the covariance matrix derived from simulated sets of parameter estimates. Thus, the approach outlined in Methods and Notation may be used to define sample sizes needed to achieve required precision.

DISCUSSION

The main purpose of this paper is to present methods and results for calculating sample sizes needed to estimate penetrance of an autosomal dominant gene with required precision from cohort, case-control, GPD, and GPDR designs. Data needed to compute such sample sizes for GPD and GPDR designs with dichotomous phenotypes are given in Tables II–V for a wide range of parameters, and methods are also presented for survival data.

In selecting an appropriate design, one should consider both the validity of the inference and the feasibility of the design. Some designs can yield seriously biased results if non-representative samples are obtained, as we discuss in the Validity section. Sample size requirements are one aspect of feasibility, but there are other practical considerations.

Feasibility

The prospective cohort study suffers the serious disadvantage that one may need to wait many years to observe disease outcomes, and large numbers of subjects may need to be screened to obtain a sufficiently large cohort of gene carriers. The historical cohort design allows one to avoid waiting for events to occur by relying on a previously constituted cohort for which stored biological materials are available for genotyping. One must examine all carriers in the cohort, review their medical records, or rely on disease registries to establish disease status of carriers. As indicated in Table I, the numbers of subjects one needs to screen to identify carriers may be formidable. If one would settle for a precision of $\pm 10\%$ instead of $\pm 5\%$ in estimating penetrance, however, the required numbers would be reduced by a factor of four. Unless consent for genotyping had been obtained earlier, there could be ethical or practical obstacles to obtaining consent from members of a large cohort, some of whom may have become lost to follow-up or died.

The case-control design will also require large sample sizes to estimate the penetrance of rare mutations (Table I), but it avoids the difficulty of waiting for health events to occur, and it affords an opportunity to obtain informed consent when cases and controls are accrued. In order to use this design, either q must be known or the probability of disease in the population must be known.

The GPD and GPDR designs usually require somewhat smaller numbers of genotypes than do cohort and case-control designs, as illustrated in Table I. Some extra work is needed, however, to determine the phenotypes of relatives, and even more effort may be required to obtain genotypes of relatives. A feasibility assessment should

take these extra costs into account. Informed consent would ordinarily be obtained from the probands in such studies, but more consideration is required concerning informed consent to contact or genotype relatives of probands.

The calculations in this paper are based on the assumption that probands are at risk of disease, as would usually be the case for an autosomal dominant disease. In fact, in the study by Struewing et al. [1997], 30% of the probands were male and therefore essentially not susceptible to breast cancer. Therefore, the contribution to the likelihood from genotyping a male proband [see equation (2.6)] is independent of the penetrance parameters that apply to women. Thus, male probands are less informative regarding penetrance than female probands, although the contribution to the likelihood from phenotyping the female relatives of male probands is still informative [see equation (2.7)]. If males are to be included as probands and are regarded as non-susceptible, sample sizes can be estimated by using the methods in this paper with $\epsilon_0 = P(g_0 = 1)$ in equation (2.9).

Validity

The internal validity of a cohort study depends on identifying all initial cohort members, correctly genotyping them, and, for purposes of estimating the penetrance of carriers, obtaining a complete and accurate assessment of the disease status of all carriers. For the results to be generalizable, the initial cohort screened should be typical of the general population, unless it can be assumed that other factors apart from genotype have no influence on the probability of disease. Two problems need to be avoided. First, all carriers or a representative sample of the carriers in the cohort must be identified. If, instead, one tended to preferentially genotype subjects with disease, then carriers without disease would be under-represented, leading to overestimates of penetrance. Second, follow-up of all carriers should be complete. If non-diseased carriers tended to be lost to follow-up more than diseased carriers, penetrance could again be overestimated. These two problems can be avoided by properly defining the cohort and using active follow-up procedures that apply equally to diseased and non-diseased subjects.

One issue that is hard to control in prospective cohort studies is the possibility that carriers will take preventive action, such as women who take tamoxifen to reduce breast cancer risk. The penetrance estimate will apply under these conditions of preventive care. If one wishes to understand the natural history of the disease, historical cohort studies or other designs would be preferable. (We thank Mr. Laurence Freedman for pointing this out.)

The validity of penetrance estimates for the case-control design depends principally on the ability to obtain representative samples of cases and controls. Ideally, cases and controls are obtained from a population-based design that allows for probability sampling. If cases with a strong family history or from families known to carry the mutation have higher participation rates in the study than other cases or controls selected for inclusion, penetrance will be overestimated. Likewise, if controls tend to refuse to participate if they have a strong family history or if the mutation is known to be segregating in their family, penetrance will be overestimated. If participation rates depend only on family history, and not on known mutations in the family, however, these biases are likely to be modest, as we illustrate in connection with Table VI.

TABLE VI. Distribution of Families by the Phenotypes (Y) and Genotypes (g) of Proband Daughters (g_0, Y_0) and Their Mothers (g_1, Y_1)*

Y_1	1	1	1	1	0	0	0	0
g_0	1	1	0	0	1	1	0	0
Y_0	1	0	1	0	1	0	1	0
$g_1 = 0$	302	26	9,901	89,112	2,714	236	89,112	802,006
$g_1 = 1$	2,802	244	302	2,714	244	21	26	236

*This distribution of 999,998 families is based on the parameters in the legend to Table I, modelled on the data in Claus et al. [1991].

The GPD and GPDR designs are subject to much more severe biases than cohort or case-control designs if the probands who volunteer tend to be those who have families with affected relatives. Table VI describes the distribution of (Y_1, g_1, Y_0, g_0) in a population of 999,998 families consisting of mothers with (Y_1, g_1) and proband daughters with (Y_0, g_0) . The distribution corresponds exactly to the parameters and assumptions in Table I, the motivating example. In particular, in this population, $\phi_1 = P(Y_0 = 1 | g_0 = 1) = 6062 / (6062 + 527) = .92002$, which equals .92, apart from rounding. Likewise $\phi_2 = P(Y_1 = 1 | g_1 = 1) = 6062 / (6062 + 527) = .92002$.

Suppose the only daughters in this study who volunteer as probands are those with affected mothers ($Y_1 = 1$). Among such daughters, 3,104 have $g_0 = 1, Y_0 = 1$, and 270 have $g_0 = 1, Y_0 = 0$. Thus a cohort study would yield the correct penetrance estimate $3,104 / (270 + 3104) = .9200$.

A case-control study conducted in this same subpopulation with $Y_1 = 1$ would yield the estimate $\hat{P}(g_0 = 1 | Y_0 = 1) = .23326$, rather than the correct value of 0.0575 for the entire population. Likewise $\hat{P}(g_0 = 1 | Y_0 = 0) = .002932$, rather than the correct value, 0.00059, and $\hat{P}(Y_0 = 1) = .1262$, rather than the correct value 0.1054. Nonetheless, substituting each of these incorrect values in equation (2.2), we estimate ϕ_1 as 0.920, the correct value! Even if we use the correct population probability $P(Y = 1) = 0.1054$ in equation (2.2), we estimate ϕ_1 as .904, which is close to correct.

To understand why cohort and case-control designs are not susceptible to bias from participation rates that depend only on family history, consider the cohort design. Under the conditional independence assumption that Y_0 is independent of Y_1 given g_0 , the cohort obtained by sampling with probabilities determined by Y_1 will still yield unbiased estimates of $P(Y_0 | g_0)$. This intuition can be made rigorous by defining the indicator $U = 1$ if a person is included in the cohort and 0 otherwise, where $P(U = 1 | Y_0, Y_1, g_0) = P(U = 1 | Y_1)$ depends only on Y_1 . Then, for members of the selected cohort,

$$\begin{aligned}
 P(U = 1, Y_0, g_0) &= \sum_{Y_1} P(Y_1) P(U = 1 | Y_1) P(g_0 | Y_1, U = 1) P(Y_0 | g_0, Y_1, U = 1) \\
 &= \sum_{Y_1} P(Y_1) P(U = 1 | Y_1) P(g_0 | Y_1, U = 1) P(Y_0 | g_0) \\
 &= K(g_0) P(Y_0 | g_0),
 \end{aligned}$$

where K is a positive function depending only on g_0 . Thus $P(Y_0 | g_0, U = 1) = P(U = 1, Y_0, g_0) / P(g_0, U = 1) = P(Y_0 | g_0)$, which reduces to ϕ_1 for $g_0 = Y_0 = 1$. Because in expectation the case-control design yields the same estimates as the cohort design, it also yields unbiased estimates of ϕ_1 and ϕ_0 . As we illustrate below, however, if the

selection probability $P(U = 1|Y_0, Y_1)$ depends both on Y_0 and Y_1 , which we term “differential non-response bias,” a modest bias in $\hat{\phi}_1$ is introduced with cohort and case-control designs.

In contrast, both the GPD and GPDR designs lead to serious overestimates of ϕ_1 in this example. In fact, for the GPD design with q unknown, $\hat{\phi}_1 = .99998$ and $\hat{\phi}_0 = .99908$, far from the correct values $\phi_1 = .92$ and $\phi_0 = .10$. Likewise the GPDR design with all mothers genotyped yields $\hat{\phi}_1 = .99999$ and $\hat{\phi}_0 = .99978$. The estimates of ϕ_1 and ϕ_0 from the GPD and GPDR designs are exactly 1.0 if one discards the contribution to the likelihood from $P(g_0|Y_0)$. This is admittedly an extreme example, but it illustrates that the GPD and GPDR designs are much more susceptible to biased sampling based on family history than are cohort or case-control designs.

A modest bias can be induced in cohort and case-control studies, as well as in GPD and GPDR studies, if participation depends differentially on family history for probands with and without disease. For example, suppose that all daughters with disease ($Y_0 = 1$) whose mothers are also affected ($Y_1 = 1$) participate, but only 50% of other daughters participate. Then, from Table VI the numbers of daughters with $(g_0, Y_0) = (1,1), (1,0), (0,1),$ and $(0,0)$ are, respectively, 4,583, 263.5, 54,772, and 447,034. In this selected population, the estimate of ϕ_1 from a cohort study is $4,583/(4,583 + 263.5) = .946$. For a case-control study, the estimate of ϕ_1 is, from equation (2.2), .946, where $P(Y_0 = 1)$ is estimated as $59,355/506,652 = .1172$ in this selected population. If the original population value $P(Y_0 = 1) = .1054$ is used instead, we estimate $\hat{\phi}_1 = .939$ from equation (2.2). Thus, differential non-response on the basis of family history and disease status of the proband induces a modest bias in the estimate of ϕ_1 for the cohort and case-control designs. Differential non-response induces a similar degree of bias for GPD and GPDR designs in this case. If the allele frequency is unknown, the corresponding estimates of ϕ_1 from the GPD and GPDR designs are, respectively, .956 and .945. As explained previously, non-differential non-response based on family history induces no bias in cohort or case-control estimates of ϕ_1 , provided the conditional independence assumption holds, whereas the GPD and GPDR designs are highly susceptible to bias in this situation.

The information obtained from the proband on the medical history of relatives in the GPD design may be less reliable than medical history obtained directly from participants in a case-control study, particularly if one is seeking information on age at disease onset.

We have validated by simulation in a few cases that the sample sizes calculated for GPD designs yield the promised precision for estimating ϕ_1 . These calculations are based on asymptotic methods. We have noticed that if one uses much smaller samples, the distribution of $\hat{\phi}_1$ may be skewed or have considerable mass on a boundary like $\hat{\phi}_1 = 1$, calling into question the validity of standard asymptotic inference. Further research would be needed, for example, to define appropriate interval estimates of ϕ_1 for small samples and to develop sample size estimates for small studies with poor precision.

A reviewer pointed out that Hardy-Weinberg equilibrium (HWE) would not be satisfied if the population consisted of strata with varying allele frequencies and if members of such strata only mated with others within their stratum. If stratum mem-

bership can be determined, a stratified analysis that assumed HWE within each stratum could be performed, and, under the assumption that ϕ_0 and ϕ_1 are constant across strata, combined stratified estimates of ϕ_0 and ϕ_1 could be obtained as weighted averages of the stratum-specific estimates. Maximum likelihood based on the product of stratum-specific likelihoods could also be used. Sample size calculations could be derived from this modified likelihood.

If stratum membership is unknown, the model based on HWE is incorrect for the GPD and GPDR analyses. One can determine the biases that result from using the methods in this paper when, in fact, the population is stratified. As an example, consider two strata constituting 20 and 80% of the population, respectively, and with respective allele frequencies 0.0150890903 and 0.0003744514. These allele frequencies were chosen so that the carrier frequency in the whole population is $P(g = 1) = 1 - (1 - 0.0033)^2 = 0.00658911$, just as in Table I. Examination of the score equations shows that in large samples with $\rho = 0.1$ the estimate of ϕ_1 converges to 0.920010 if q is assumed known and to 0.920011 if q is assumed unknown. There is, therefore, a very small asymptotic bias compared to the true penetrance, 0.92, that holds for a population with this carrier frequency under HWE. Respective estimates of ϕ_0 converge to 0.100014 and 0.00014, compared to 0.10, and the estimates of q in the three-parameter model converges to 0.0033. Sample size calculations needed to obtain a precision of $\pm 5\%$, as in Table I, require the use of a “sandwich estimator” for the variance, because the variance of the score does not equal the expected derivative of the score when the model is misspecified. The sample sizes required based on the sandwich procedure are 16,061 if q is assumed known and 16,075 if q is assumed unknown, which hardly differ from the corresponding values 16,068 and 16,080 in Table I. These calculations indicate that stratification in the presence of small allele frequencies usually has only a minor impact on estimates of penetrance and required sample sizes for the GPD.

Stratification of this type has no impact on inference from cohort studies, and even the sample size requirements for cohort studies will be unaffected so long as the carrier frequency in the whole population remains fixed and penetrances are constant over strata. Likewise, inference and sample size calculations for population-based case-control studies will be unaffected provided the carrier frequency in the whole population is fixed, penetrances are constant across strata, and representative samples of cases and control are obtained from the entire population. Letting $P(j)$ denote the proportion of the population in stratum j , we calculate

$$\begin{aligned} P(g = 1|Y) &= \sum_j P(j)P(g = 1|j)\phi_1 / \sum_j P(j)\{P(g = 1|j)\phi_1 + P(g = 0|j)\phi_0\} \\ &= \phi_1 P(g = 1) / \{\phi_1 P(g = 1) + \phi_0 P(g = 0)\}, \end{aligned}$$

which is the same for any stratification scheme that preserves the carrier frequency, $P(g = 1)$.

We have assumed that phenotypes within a family are conditionally independent, given carrier status, for our purposes of study design. One should, of course, consider the possibility of residual familial correlation during the analysis of such a study [see, e.g., Li and Thompson, 1997].

ACKNOWLEDGMENTS

Professor Carroll's research was supported by a grant from the National Cancer Institute (CA-57030) and partially completed during a visit to the Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute. We thank Dr. Jeff Struewing and his colleagues for making available data on the ages of participants in the study by Struewing et al. [1997] and Mrs. Holly Brown for typing the manuscript.

REFERENCES

- Claus EB, Risch N, Thompson WD (1991): Genetic analysis of breast cancer in the Cancer and Steroid Hormone Study. *Am J Hum Genet* 48:232–242.
- Cornfield J (1951): A method for estimating comparative rates from clinical data: Applications to cancer of the lung, breast and cervix. *J Natl Cancer Inst* 11:1269–1275.
- Easton DF, Ford D, Bishop T, and the Breast Cancer Linkage Consortium (1995): Breast and ovarian cancer incidence in BRCA1-mutation carriers. *Am J Hum Genet* 56:265–271.
- Li H, Thompson E (1997): Semiparametric estimation of major gene and family-specific random effects for age of onset. *Biometrics* 53:282–293.
- Prentice RL, Kalbfleisch JD, Peterson AV, et al. (1978): The analysis of failure times in the presence of competing risks. *Biometrics* 34:541–554.
- Rao CR (1965): "Linear Statistical Inference and Its Applications," 2nd ed. New York: John Wiley.
- Struewing JP, Hartge P, Wacholder S, et al. (1997): The risk of cancer associated with specific mutations of BRCA1 and BRCA2 among Ashkenazi Jews. *N Engl J Med* 336:1401–1408.
- Wacholder S, Hartge P, Struewing JP, et al.: The kin-cohort study for estimating penetrance. *Am J Epidemiol* (in press).
- Whittemore AS, Gong G, Itnyre J (1997): Prevalence and contribution of BRCA1 mutations in breast cancer and ovarian cancer: Results from three U.S. population-based case-control studies of ovarian cancer. *Am J Hum Genet* 60:496–504.
- Wingo PA, Ory HW, Layde PM, et al. (1988): The evaluation of the data collection process for a multicenter, population-based, case-control design. *Am J Epidemiol* 128:206–207.

APPENDIX: LIKELIHOODS AND INFORMATION FOR GPD AND GPDR DESIGNS

GPD Design

Suppose there are N proband cases ($Y_0 = 1$) and M proband controls ($Y_0 = 0$). Let N_{ijk} be the number of families of case probands with $g_0 = i$, $Y_1 = j$, $Y_2 = k$, and define M_{ijk} similarly for families of control probands. Let $p_{jk} = P(g_1 = j, g_2 = k | g_0 = 1)$ and $w_{jk} = P(g_1 = j, g_2 = k | g_0 = 0)$, and let $\epsilon_1 = P(g_0 = 1 | Y_0 = 1)$, and $\epsilon_0 = P(g_0 = 1 | Y_0 = 0)$, as in equations (2.8) and (2.9). Finally, let Z be the number of proband cases who are carriers ($g_0 = 1$), and X be the number of proband controls who are carriers ($g_0 = 1$). Note that p_{jk} and w_{jk} are functions of the allele frequency q .

The total likelihood for the $N + M$ families is $L = L$ (case probands) \times L (control probands). Here L (case probands) =

$$\begin{aligned}
& \varepsilon_1^Z (1 - \varepsilon_1)^{N-Z} \{\phi_1^2 p_{11} + \phi_1 \phi_0 p_{10} + \phi_0 \phi_1 p_{01} + \phi_0^2 p_{00}\}^{N_{111}} \\
& \times \{\phi_1 (1 - \phi_1) p_{11} + \phi_1 (1 - \phi_0) p_{10} + \phi_0 (1 - \phi_1) p_{01} + \phi_0 (1 - \phi_0) p_{00}\}^{N_{110}} \\
& \times \{(1 - \phi_1) \phi_1 p_{11} + (1 - \phi_1) (\phi_0) p_{10} + (1 - \phi_0) \phi_1 p_{01} + (1 - \phi_0) \phi_0 p_{00}\}^{N_{101}} \\
& \times \{(1 - \phi_1)^2 p_{11} + (1 - \phi_1) (1 - \phi_0) p_{10} + (1 - \phi_0) (1 - \phi_1) p_{01} + (1 - \phi_0)^2 p_{00}\}^{N_{100}} \\
& \times \{\phi_1^2 w_{11} + \phi_1 \phi_0 w_{10} + \phi_0 \phi_1 w_{01} + \phi_0^2 w_{00}\}^{N_{111}} \\
& \times \{\phi_1 (1 - \phi_1) w_{11} + \phi_1 (1 - \phi_0) w_{10} + \phi_0 (1 - \phi_0) w_{01} + \phi_0 (1 - \phi_0) w_{00}\}^{N_{010}} \\
& \times \{(1 - \phi_1) \phi_1 w_{11} + (1 - \phi_1) (\phi_0) w_{10} + (1 - \phi_0) \phi_1 w_{01} + (1 - \phi_0) \phi_0 w_{00}\}^{N_{001}} \\
& \times \{(1 - \phi_1)^2 w_{11} + (1 - \phi_1) (1 - \phi_0) w_{10} + (1 - \phi_0) (1 - \phi_1) w_{01} + (1 - \phi_0)^2 w_{00}\}^{N_{000}}.
\end{aligned} \tag{A1}$$

The expression for L (control probands) is identical to equation (A1) except that X, M, M_{ijk} , and ε_0 , replace, respectively, Z, N, N_{ijk} , and ε_1 .

The information matrix for a study of a single case proband family is obtained by setting N = 1 and by numerically differentiating the logarithm of -L(case proband) twice with respect to ϕ_1 , ϕ_0 , and q (unless q is assumed known), while Z and $\{N_{ijk}\}$ are set at their expected values. For example, $EZ = \varepsilon_1$ and $E(N_{111}) = \phi_1^2 p_{11} + \phi_1 \phi_0 p_{10} + \phi_0 \phi_1 p_{01} + \phi_0^2 p_{00}$. These expectations are regarded as fixed constants and not as functions of ϕ_1 , ϕ_0 , and q for purposes of differentiation. The information matrix for a control proband family is found similarly.

The information from families with a single relative with phenotype Y_1 is obtained as in equation (A1). Letting N_{ij} denote the number of families of case probands with $g_0 = i$ and $Y_1 = j$, M_{ij} denote the number of such families of control probands, $p_j = P(g_1 = j | g_0 = 1)$, and $w_j = P(g_1 = j | g_0 = 0)$, we express the likelihood for N families with case probands as

$$\begin{aligned}
& \varepsilon_1^Z (1 - \varepsilon_1)^{N-Z} \{\phi_1 p_1 + \phi_0 p_0\}^{N_{11}} \{1 - \phi_1\} p_1 + (1 - \phi_0) p_0\}^{N_{10}} \\
& \times \{\phi_1 w_1 + \phi_0 w_0\}^{N_{01}} \{(1 - \phi_1) w_1 + (1 - \phi_0) w_0\}^{N_{00}}
\end{aligned} \tag{A2}$$

where Z of the N case probands were carriers. The likelihood for M control proband families is given by equation (A2) with X, M, M_{ij} , and ε_0 replacing Z, N, N_{ij} , and ε_1 , respectively. Here X is the number of control probands who are carriers. The information matrix is obtained by numerical differentiation as described for the case of two relatives.

The expressions (A1) and (A2) apply for any type of relative of the proband provided the conditional distributions p_{jk} and w_{jk} or p_j and w_j , are calculated appropriately.

GPDR Design

First, we consider families with one relative in which both the proband and relative will be genotyped. Suppose there are N families with case probands. Let R_{ijk} be the number of families with $g_0 = i$, $g_1 = j$, $Y_1 = k$. The likelihood for these families is

$$\begin{aligned}
 L(\text{case proband}) &= (\varepsilon_1 p_1 \phi_1)^{R_{111}} \{\varepsilon_1 p_1 (1 - \phi_1)\}^{R_{110}} \{\varepsilon_1 p_1 \phi_0\}^{R_{101}} \{\varepsilon_1 p_0 \phi_0\}^{R_{100}} \\
 &\times \{(1 - \varepsilon_1) w_1 \phi_1\}^{R_{011}} \{(1 - \varepsilon_1) w_1 (1 - \phi_1)\}^{R_{010}} \{(1 - \varepsilon_1) w_0 \phi_0\}^{R_{001}} \{(1 - \varepsilon_1) w_0 (1 - \theta_0)\}^{R_{000}}.
 \end{aligned}
 \tag{A3}$$

The likelihood for families with control probands, $L(\text{control proband})$, is identical except that N is replaced by M , ε_1 is replaced by ε_0 , and R_{ijk} now corresponds to families with control probands. The information matrix is obtained by numerical differentiation of minus the log-likelihood with random variables R_{ijk} , regarded as constants, set at their expected values. For example, to compute the information from a single case proband family we set $N = 1$, and, corresponding to equation (A3), we set $E(R_{111}) = \varepsilon_1 p_1 \phi_1$, $E(R_{110}) = \varepsilon_1 p_1 (1 - \phi_1)$, . . . , $E(R_{000}) = (1 - \varepsilon_1) w_0 (1 - \phi_0)$.

If we only genotype the relative when the proband is a gene carrier ($g_0 = 1$), the likelihood is modified. For case probands, the likelihood becomes

$$\begin{aligned}
 &(\varepsilon_1 p_1 \phi_1)^{R_{111}} \{\varepsilon_1 p_1 (1 - \phi_1)\}^{R_{110}} (\varepsilon_1 p_0 \phi_0)^{R_{101}} \{\varepsilon_1 p_0 (1 - \phi_0)\}^{R_{100}} \\
 &\times [(1 - \varepsilon_1) \{w_1 \phi_1 + w_0 \phi_0\}]^{R_{0+1}} [(1 - \varepsilon_1) \{w_1 (1 - \phi_1) + w_0 (1 - \theta_1) + w_0 (1 - \phi_0)\}]^{R_{0+0}},
 \end{aligned}
 \tag{A4}$$

where $R_{0+1} = R_{011} + R_{001}$ and $R_{0+0} = R_{010} + R_{000}$. The information matrix is calculated as before with $E(R_{0+1}) = E(R_{011}) + E(R_{001})$ and $E(R_{0+0}) = E(R_{010}) + E(R_{000})$.