Am. J. Hum. Genet. 52:1129–1138, 1993

Estimation of the Incidence of a Rare Genetic Disease through a Two-Tier Mutation Survey

Ranajit Chakraborty,* M. R. Srinivasan,* and Salmo Raskin†,‡

*Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, Houston; †Department of Pediatrics, Vanderbilt University School of Medicine, Nashville; and ‡Universidade Federal do Parana, Curitiba, Brazil

Summary

Recent attempts to detect mutations involving single base changes or small deletions that are specific to genetic diseases provide an opportunity to develop a two-tier mutation-screening program through which incidence of rare genetic disorders and gene carriers may be precisely estimated. A two-tier survey consists of mutation screening in a sample of patients with specific genetic disorders and in a second sample of newborns from the same population in which mutation frequency is evaluated. We provide the statistical basis for evaluating the incidence of affected and gene carriers in such two-tier mutation-screening surveys, from which the precision of the estimates is derived. Sample-size requirements of such two-tier mutation-screening surveys are evaluated. Considering examples of cystic fibrosis (CF) and medium-chain acyl-CoA dehydrogenase deficiency (MCAD), the two most frequent autosomal recessive disease in Caucasian populations and the two most frequent mutations (ΔF508 and G985) that occur on these disease allele–bearing chromosomes, we show that, with 50–100 patients and a 20-fold larger sample of newborns screened for these mutations, the incidence of such diseases and their gene carriers in a population may be quite reliably estimated. The theory developed here is also applicable to rare autosomal dominant diseases for which disease-specific mutations are found.

Introduction

For a large fraction of the 5,000 Mendelian diseases in humans known thus far (McKusick 1992), the genes are identified, and specific mutations causing the disease have been detected. The incidence of a rare disease is difficult to assess precisely, because other causes may obscure the detection of affected individuals before their death. Molecular techniques for detection of disease-specific mutations have the potential to ameliorate this problem, although screening for specific mutations does not generally detect all affected individuals. For example, recently the cystic fibrosis (CF) gene has been identified (Kerem et al. 1989), and it is known that, worldwide, nearly 70% of the CF allele(s) carry a deletion mutation (ΔF508) at codon 508 of the CF transmembrane conductance regulator (CFTR) protein (Cystic Fibrosis Genetic Analysis Consortium 1990). This provides the opportunity to screen for this mutation in newborns, from which the disease incidence in the population may be predicted. Similar screening for disease-specific mutations (base-pair changes or small deletion) may also apply for estimation of incidences of α1-antitrypsin deficiency (Newton et al. 1989), sickle cell anemia (Wu et al. 1989), phenylketonuria (Sommer et al. 1989), apolipoprotein E (Wenham et al. 1991), β-thalassemia (Old et al. 1990), medium-chain acyl-CoA dehydrogenase deficiency (MCAD; Yokota et al. 1991), etc.

In all such cases, it is important to know the frequencies of disease-specific mutations in a general population, as well as the fraction of disease allele(s)–bearing chromosomes that carry the disease-specific mutations. For some of the diseases mentioned above, screening surveys have accomplished this by comparing the frequencies of disease-specific mutations on disease gene-

Received August 25, 1992; final revision received February 3, 1993.

Address for correspondence and reprints: Dr. Ranajit Chakraborty, Center for Demographic and Population Genetics, University of Texas Graduate School of Biomedical Sciences, P.O. Box 20334, Houston, TX 77225.

0002-9297/93/5206-0014$02.00

bearing and normal chromosomes (e.g., see Cystic Fibrosis Genetic Analysis Consortium 1990; Romeo and Devoto 1990). A two-tier sampling procedure should also provide such information, when (1) one of the samples relates to mutation screening among a patient population and (2) in the other mutation screening is conducted in the general population (newborns). The purpose of this work is to determine the sample-size requirement and to evaluate the precision of the estimates for such two-tier screening programs. In addition, we derive the confidence intervals for each of these parameters from which heterogeneity of incidences in different populations may be evaluated. The theory is illustrated with data on CF and MCAD, by considering screening for the ΔF508 and G985 mutations, which are specific for these diseases, respectively.

Theory

Consider an autosomal recessive genetic disease and assume that the mutation (say, M) for which the screening survey is designed occurs only on the disease allele–bearing chromosome. Let d be the true frequency of the disease allele(s) in a population, so that d^2 is the frequency of affected individuals in the population, and $2d(1-d)$ is the frequency of carriers in the population.

Estimates and Their Standard Errors

In a sample of m patients, let m_1 individuals be the number of homozygotes (MM) for the mutation, m_2 the number of heterozygotes ($M\bar{M}$), and m_3 the number of homozygotes ($\bar{M}\bar{M}$) for the absence of the mutation. The gene-count estimator of r, the proportion of the disease alleles carrying the mutation M, is $\hat{r} = (2m_1 + m_2)/2m$, with a variance $V(\hat{r}) = r(1-r)/2m$.

Similarly, in screening n newborns (whose disease status is still unknown) for mutation M, let n_1 individuals be the carriers of the mutation M. The rarity of disease in populations will generally yield no mutant homozygote in a random sample of newborns from the population. This screening survey results in a gene-count estimator of a composite parameter, $p = d \cdot r$. The estimate of p (from allele counts) becomes $\hat{p} = n_1/2n$, with the sampling variance $V(\hat{p}) = p(1-p)/2n$.

Thus, the estimate of the disease allele frequency is given by

$$\hat{d} = \hat{p}/\hat{r}, \tag{1}$$

whose approximate sampling variance is (see Kendall 1947)

$$V(\hat{d}) \approx d^2\left(\frac{V(\hat{p})}{p^2} + \frac{V(\hat{r})}{r^2}\right). \tag{2}$$

We show below that this approximation is quite accurate for the problem that we are considering.

Under the Hardy-Weinberg assumption, the disease incidence in the population is estimated by $\hat{d}^2$, which has an approximate sampling variance of

$$V(\hat{d}^2) \approx 4d^2 \cdot V(\hat{d}). \tag{3}$$

Similarly, the carrier frequency in the population is estimated by $2\hat{d}(1-\hat{d})$, and its approximate sampling variance is

$$V[2\hat{d}(1-\hat{d})] \approx 4(1-2d)^2 \cdot V(\hat{d}). \tag{4}$$

The precision of these estimates may be judged by their coefficient of variation. The estimators from different population samples may be contrasted by using Rao's (1973) heterogeneity χ^2 analysis.

Confidence Intervals for Disease Incidence and Carrier Frequencies

The standard errors derived above cannot be directly used to obtain the confidence intervals of these parameters, because, when d is small, the sampling distribution of the estimates of d^2 and $2d(1-d)$ are not symmetric around their point estimates. Two approaches may be considered to derive confidence-interval estimates of the disease incidence, d^2, and the frequency of carriers, $2d(1-d)$, in a population.

First, since the estimators of d^2 and $2d(1-d)$ are

$$\hat{d}^2 = \left(\frac{n_1}{2m_1 + m_2} \cdot \frac{m}{n}\right)^2, \tag{8}$$

and

$$2\hat{d}(1-\hat{d}) = 2\left(\frac{n_1}{2m_1 + m_2} \cdot \frac{m}{n}\right)\left(1 - \frac{n_1}{2m_1 + m_2} \cdot \frac{m}{n}\right), \tag{9}$$

their sampling distributions are uniquely specified by those of n_1 and $(2m_1 + m_2)$, which are binomial variates with parameters $(2n,p)$ and $(2m,r)$, respectively. Further-

more, by the design of the study, n_1 and $(2m_1+m_2)$ are independently distributed. Therefore, the probability of observing the point estimates (expressions [8] and [9]) is

$$\binom{2n}{n_1} \cdot (rd)^{n_1} \cdot (1-rd)^{2n-n_1} \cdot \binom{2m}{2m_1+m_2} \cdot r^{2m_1+m_2} \cdot (1-r)^{m_2+2m_3},$$

which can be computed for all combinations of $2m_1+m_2 = 0, 1, 2, \ldots, 2m$ and $n_1 = 0, 1, 2, \ldots, 2n$, for any given value of r and d. By sorting the probabilities in ascending order of values of the estimated d^2 and $2d(1-d)$, we can obtain the lower and upper confidence bounds for both estimators for any specified sample sizes m and n.

However, as the exact enumeration of confidence bounds is tedious for large sample sizes, we may approach the confidence-interval evaluation for d^2 and $2d(1-d)$ by considering their logarithmic transformation, which makes the point estimators almost symmetrically distributed around their expectations (Chakraborty et al. 1993). The 95% confidence interval of $\hat{x} = \ell n(\hat{d}^2) = 2\ell n(\hat{d})$ is given by

$$\hat{x} \pm 1.96 \cdot \{V(\hat{x})\}^{1/2}, \qquad (10)$$

in which $V(\hat{x}) \approx 4 \cdot V(\hat{d})/d^2$, where $V(\hat{d})$ is evaluated from equation (2) by substituting the estimates for the parameters. Once the confidence interval (c_L, c_U) for x is obtained from equation (10), a reverse transformation will generate the 95% confidence interval for the disease incidence, given by (e^{c_L}, e^{c_U}). By analogy, the confidence interval for the carrier frequencies may be obtained from a transformed parameter, $y = \ell n[2d(1-d)]$, whose estimate is $\hat{y} = \ell n[2\hat{d}(1-\hat{d})]$. Its variance is

$$V(\hat{y}) \approx (1-2d)^2 \cdot V(\hat{d})/[d^2(1-d)^2], \qquad (11)$$

so that an equation of form (10) will generate the confidence interval for the incidence of gene carriers, $2d(1-d)$. Although the second approach involves approximations (asymptotic normality of the logarithmic transformed sample statistics), we show in the sequel that this simpler method is adequate when m is 25 or more and n is at least five times larger than m.

Sample-Size Requirements

The above theory can also be used to determine the minimum sample size for obtaining a given precision of the point estimates. First, suppose that we would like to estimate both d^2 and $2d(1-d)$ with coefficient of variation not exceeding a certain fraction c (generally, $c = .25$–$.5$ is an indicator of a precise estimate). From equation (3), we obtain that the sample size for estimating the disease incidence in a population n should satisfy

$$n \geq \left(\frac{1-rd}{rd}\right) \Big/ \left(\frac{c^2}{2} - \frac{1-r}{mr}\right), \qquad (12)$$

while, for estimating the incidence of disease-gene carriers (from eq. [4]), the sample size should satisfy

$$n \geq \left(\frac{1-rd}{rd}\right) \Big/ \left(\frac{2c^2(1-d)^2}{(1-2d)^2} - \frac{1-r}{mr}\right). \qquad (13)$$

Numerical calculations, shown below, suggest that, for given values of parameters r and d, with any specified sample sizes n and m, the disease-gene carrier frequency is more precisely estimated than is the disease incidence.

Similarly, sample-size requirements may also be determined from the confidence-interval estimation of d^2 and $2d(1-d)$, by fixing the ratio of upper versus lower confidence bounds, for a specified level of confidence. Suppose that we wish to determine 95% confidence bounds of d^2 such that the ratio of the upper and lower 95% bounds remains less than a specified value, say R (generally $R = 2$, 5, or even 10 may be regarded as reasonable, since d^2 is itself a small quantity). Substituting this into equation (10), we have

$$n \geq \left(\frac{1-rd}{rd}\right) \Big/ \left(\frac{(\ell n\, R)^2}{30.7328} - \frac{1-r}{mr}\right), \qquad (14)$$

and an equivalent approach for estimating $2d(1-d)$ yields

$$n \geq \left(\frac{1-rd}{rd}\right) \Big/ \left(\frac{2(1-d)^2(\ell n\, R)^2}{15.3664(1-2d)^2} - \frac{1-r}{mr}\right), \qquad (15)$$

specifying the sample-size requirements with precisions determined by their respective confidence bounds.

Applications

The theory described above may be applied to screening surveys for mutations specific to lethal autosomal recessive diseases such as CF and MCAD. While both of these diseases are more common in Caucasian populations, their precise incidences in different geographic regions are still unknown (Allan and Phelan 1985; Kolvraa et al. 1991; Yokota et al. 1991; Braekeleer and Daigneault 1992). Several specific mutations that occur on the CF- and MCAD-bearing chromosomes are known. The most common mutation for CF is the deletion of a phenylalanine residue at codon 508 (ΔF508) of the 1,480-amino-acid CFTR protein (Kerem et al. 1989). CF incidence is roughly 1/2,500 live births among Caucasians (Boat et al. 1989), yielding a gene frequency of $d = .02$. Worldwide, the ΔF508 mutation is present on approximately 70% of the CF chromosomes, giving a value of $r = .70$, although evidence of geographic variation in d and r among different Caucasian populations is abundant (Braekeleer and Daigneault 1992). Incidence of the autosomal recessive disorder of fatty-acid oxidation, MCAD, is less precisely known. Roughly 1/5,000 live births in Caucasians in England results in this disorder (Bennett et al. 1987). Sudden deaths of children account for missing occurrences of MCAD (Kolvraa et al. 1991; Yokota et al. 1991). The above incidence value gives a value of $d = .014$, although the exact gene frequency may be smaller because of the inclusion, in the general survey, of other rare inborn errors of fatty-acid oxidation (Bennett et al. 1987). Yokota et al. (1990) discovered a point mutation in the coding region of pMCAD cDNA, an A-to-G transition at position 985 (called "G985 mutation") that is specific to MCAD-bearing chromosomes. About 90% of the disease-causing alleles in diagnosed MCAD patients carry the G985 mutation (Yokota et al. 1991), giving a value of $r = .90$, even though this also may vary in different Caucasian populations.

Tables 1 and 2 present the sample-size requirements for estimating the disease incidence (table 1) and the incidence of carriers (table 2) for these two diseases. Each table gives the number of newborns (i.e., n) for which mutation (ΔF508 and/or G985) screening is to be attempted, for given values of m, the number of patients from which the value of r is determined.

Three observations may be made from these computations. First, as the specificity r of the mutations for disease-allele detection increases, the sample-size requirement is less stringent. An increase of r from .50 to .90 amounts to almost a 50% reduction in sample size n for both disease-allele frequencies ($d = .01$ and .02), irrespective of the value of m. This has important implications for the development of a two-tier mutation-screening program, since, for both CF and MCAD, other rare disease-specific mutations are known, and there are attempts to develop a rapid and reliable amplification refractory mutation system (ARMS) for screening programs (Ferrie et al. 1992) that improves the specificity (i.e., increases the value of r). For $r = 1$ (100% specificity), a single-tier screening program in newborns is sufficient, since in this case $p = d$.

Second, the contribution of the number of patients m screened to the precision of estimates of d^2 and $2d(1 - d)$ is small even when $r < 1$. In other words, an effective mutation-screening program may not require surveying a large number of patients. As long as $r > .5$, probably 100 patients (or fewer) are enough, and efforts to increase the number of newborns n can greatly improve the precision of estimates of incidences of disease and gene bearers. Third, the sample size required for estimation of the frequency of gene bearers is smaller than what would be needed for estimation of disease incidence for a fixed level of precision. Consequently, when the sample-size requirement is established by considering the precision of the estimated disease incidence, the precision of the estimate of gene carriers will be comparatively much higher. This has an important implication for applying this theory to other genetic diseases, which will be discussed later.

Since tables 1 and 2 establish that the values of m, the number of patients to be screened for mutations, have relatively less impact on the precision of the estimates of the incidences of disease and gene bearers, in figure 1 we present the upper and lower limits of 95% confidence intervals for these parameters by exact evaluation of the distributions of the estimates of d^2 and $2d(1-d)$, shown in equations (8) and (9). For these computations we varied n, the number of newborns screened, as a multiple (k) of m, so that $n = km$. All parameter values chosen for these computations are applicable to either CF and ΔF508 mutation screening, or MCAD and G985 mutation screening.

These results indicate that, irrespective of the value of m, not much improvement in the precision of the incidences can be made by increasing n beyond a factor (k) of 20, so that mutation screening for these diseases may be effectively conducted with 50–100 patients and

Table 1

Sample-Size Requirement for Estimating Incidence d^2 of an Autosomal Recessive Disease

	Value of n for					
	d = .01			d = .02		
	Coefficient of Variation of d^2 Fixed at					
r and m	1/4	1/3	1/2	1/4	1/3	1/2
.50:						
25	. . .	13,772	2,342	. . .	6,852	1,165
50	17,689	5,777	1,895	8,800	2,873	943
100	9,365	4,477	1,730	4,659	2,228	861
200	7,581	4,024	1,658	3,772	2,002	825
.75:						
25	7,386	3,219	1,186	3,666	1,598	589
50	5,384	2,770	1,119	2,672	1,375	555
100	4,741	2,589	1,088	2,353	1,285	540
200	4,474	2,508	1,073	2,220	1,245	533
.90:						
25	4,108	2,202	914	2,036	1,091	453
50	3,794	2,109	897	1,880	1,045	445
100	3,654	2,065	889	1,811	1,023	441
200	3,588	2,044	885	1,778	1,013	439
	Ratio of 95% Confidence Limits of d^2 Fixed at					
	2	5	10	2	5	10
.50:						
25	. . .	4,494	1,502	. . .	2,236	748
50	. . .	3,096	1,305	. . .	1,540	650
100	35,327	2,679	1,225	17,575	1,333	610
200	18,715	2,510	1,188	9,311	1,249	591
.75:						
25	57,539	1,866	832	28,552	926	413
50	14,759	1,705	798	7,324	846	396
100	10,759	1,635	783	5,339	812	389
200	9,475	1,602	775	4,702	795	385
.90:						
25	9,842	1,380	656	4,876	684	323
50	8,211	1,342	647	4,068	665	321
100	7,583	1,323	643	3,757	656	319
200	7,303	1,316	641	3,619	652	318

NOTE.—An ellipsis (. . .) indicates that no sample size will meet the prescribed precision.

1,000–2,000 newborns, in those populations where the disease is not much rarer than 1/5,000.

Since in all of these computations we used the approximations suggested in equations (10)–(15), in table 3 we show the effects of these approximations. For given parameter values r and d and sample sizes m and n, we evaluated the 95% confidence-interval estimates of the incidences of disease and gene carriers for rare recessive disorders, by using the exact compound binomial distribution of the estimates given in equations (8) and (9) to compare them with the approximations.

These computations indicate that, for parameter val-

Table 2

Sample-Size Requirement for Estimating Incidence of Recessive Disease Gene Carriers, $2d(1-d)$

	Value of n for					
	d = .01			d = .02		
	Coefficient of Variation of $2d(1-d)$ Fixed at					
r and m	1/4	1/3	1/2	1/4	1/3	1/2
.50:						
25	2,273	1,092	424	1,097	530	206
50	1,851	984	406	898	479	198
100	1,693	938	398	824	457	194
200	1,624	916	394	791	446	192
.75:						
25	1,159	634	267	562	308	130
50	1,095	614	263	532	299	128
100	1,066	605	262	518	294	127
200	1,052	600	261	511	292	127
.90:						
25	895	506	218	434	246	106
50	879	501	217	427	243	106
100	871	498	217	423	242	105
200	867	497	216	421	241	105
	Ratio of 95% Confidence Limits of $2d(1-d)$ Fixed at					
	2	5	10	2	5	10
.50:						
25	8,356	655	300	3,934	318	146
50	4,542	615	291	2,192	299	142
100	3,698	596	287	1,795	290	140
200	3,384	587	285	1,646	286	139
.75:						
25	3,622	401	192	1,267	195	93
50	2,316	393	190	1,123	191	93
100	2,188	389	189	1,062	189	92
200	2,130	387	189	1,035	188	92
.90:						
25	1,855	325	158	899	158	77
50	1,788	323	157	867	157	77
100	1,756	322	157	852	156	76
200	1,741	321	157	845	156	76

ues applicable for CF (d = .02, r = .70) and MCAD (d = .01, r = .90), even when the sample sizes are as low as m = 25 and n = 250, the approximations used in this work yield confidence intervals close to the exact ones. The improvement in approximation increases with sample sizes, more so with an increase of n. The approximate confidence intervals are wider than the exact ones, suggesting that use of the simple approximations (eqs. [10]–[15]) should in fact have a larger level of confidence (>95%) than indicated.

Discussion and Conclusion

The theory developed here depends on the assumption that the population represented by the series of

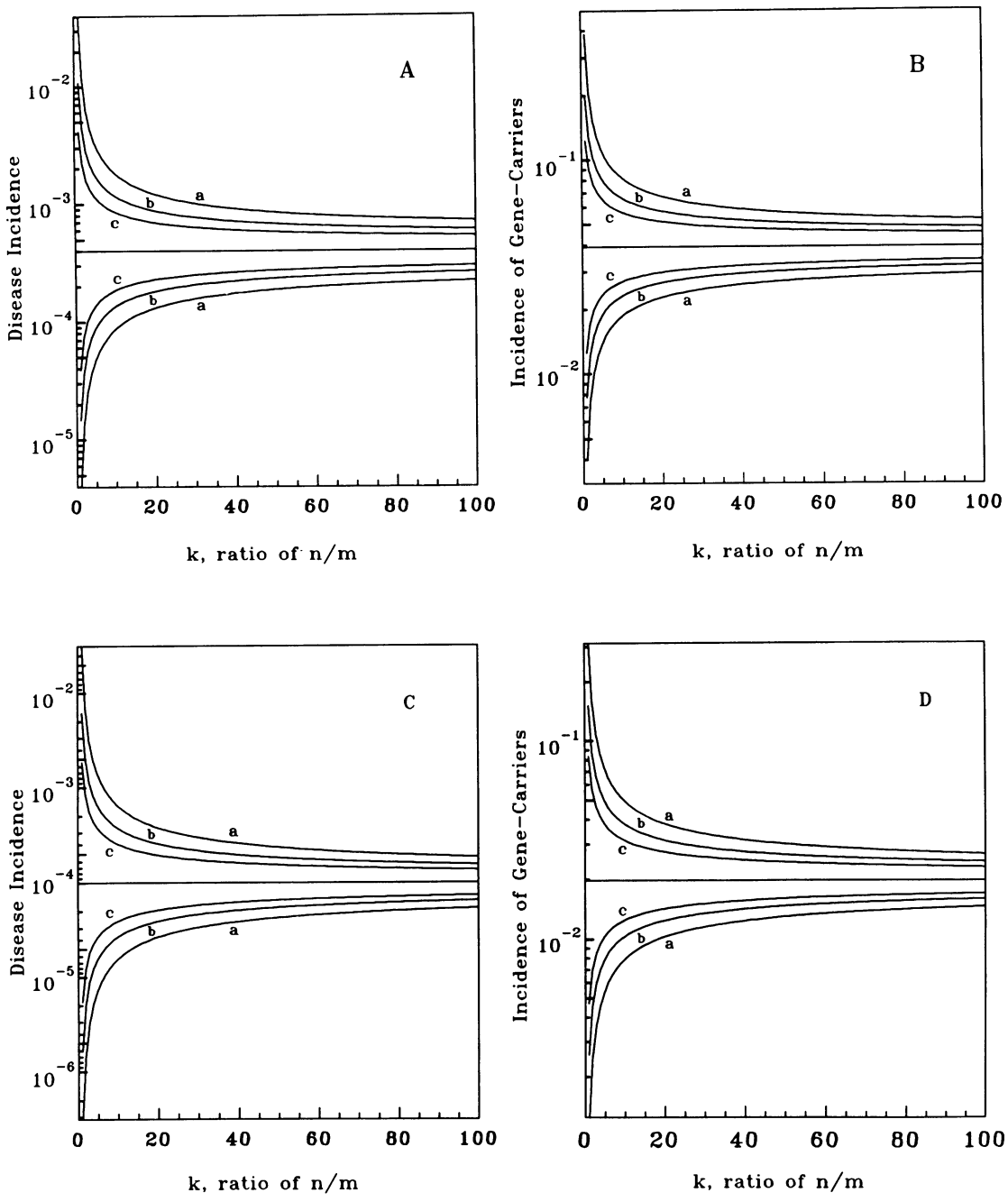

Figure 1 95% Confidence intervals for the incidences of recessive diseases (panels *A* and *C*) and gene carriers (panels *B* and *D*), through a two-tier mutation-screening survey. In each panel the three curves marked are for three values of m, the number of patients screened: $m = 25$ (*a*), $m = 50$ (*b*), and $m = 100$ (*c*). The number of newborns screened, n, is determined by the factor k; i.e., $n = km$. The computations represent parameters $d = .02$ and $r = .70$ (panels *A* and *B*) and $d = .01$ and $r = .90$ (panels *C* and *D*), applicable for autosomal recessive diseases CF and MCAD, respectively. The horizontal lines represent the true values of incidences of affected and gene carriers for the chosen parameters d and r.

Table 3

Comparisons of Exact and Approximate 95% Confidence-Interval Estimates of Incidences of Recessive Disease and Gene Carriers through a Two-Tier Mutation-Screening Survey

m AND n	CONFIDENCE BOUND	95% CONFIDENCE LIMITS OF INCIDENCES OF			
		Diseases		Gene Carriers	
		Exact	Approximate	Exact	Approximate
$d = .01, r = .90$:					
25:					
250	Lower	1.66×10^{-5}	1.58×10^{-5}	7.97×10^{-3}	7.93×10^{-3}
	Upper	6.28×10^{-4}	6.35×10^{-4}	4.89×10^{-2}	4.94×10^{-2}
500	Lower	2.76×10^{-5}	2.69×10^{-5}	1.06×10^{-3}	1.03×10^{-2}
	Upper	3.68×10^{-4}	3.72×10^{-4}	3.75×10^{-2}	3.79×10^{-2}
1,000	Lower	3.94×10^{-5}	3.91×10^{-5}	1.25×10^{-2}	1.24×10^{-2}
	Upper	2.53×10^{-4}	2.56×10^{-4}	3.13×10^{-2}	3.15×10^{-2}
50:					
250	Lower	1.67×10^{-5}	1.58×10^{-5}	7.98×10^{-3}	7.95×10^{-3}
	Upper	6.26×10^{-4}	6.32×10^{-4}	4.90×10^{-2}	4.93×10^{-2}
500	Lower	2.73×10^{-5}	2.71×10^{-5}	1.06×10^{-2}	1.04×10^{-2}
	Upper	3.65×10^{-4}	3.70×10^{-4}	3.76×10^{-2}	3.78×10^{-2}
1,000	Lower	3.96×10^{-5}	3.95×10^{-5}	1.26×10^{-2}	1.25×10^{-2}
	Upper	2.52×10^{-4}	2.53×10^{-4}	3.13×10^{-2}	3.14×10^{-2}
$d = .02, r = .70$:					
25:					
250	Lower	8.84×10^{-5}	8.79×10^{-5}	1.91×10^{-2}	1.87×10^{-2}
	Upper	1.78×10^{-3}	1.82×10^{-3}	8.18×10^{-2}	8.23×10^{-2}
500	Lower	1.36×10^{-4}	1.33×10^{-4}	2.32×10^{-2}	2.29×10^{-2}
	Upper	1.17×10^{-3}	1.20×10^{-3}	6.69×10^{-2}	6.72×10^{-2}
1,000	Lower	1.78×10^{-4}	1.76×10^{-4}	2.62×10^{-2}	2.62×10^{-2}
	Upper	9.07×10^{-4}	9.08×10^{-4}	5.85×10^{-2}	5.86×10^{-2}
50:					
250	Lower	9.03×10^{-5}	8.98×10^{-5}	1.93×10^{-2}	1.89×10^{-2}
	Upper	1.73×10^{-4}	1.78×10^{-3}	8.20×10^{-2}	8.15×10^{-2}
500	Lower	1.39×10^{-4}	1.37×10^{-4}	2.34×10^{-2}	2.32×10^{-2}
	Upper	1.16×10^{-3}	1.17×10^{-3}	6.61×10^{-2}	6.63×10^{-2}
1,000	Lower	1.85×10^{-4}	1.84×10^{-4}	2.68×10^{-2}	2.68×10^{-2}
	Upper	8.72×10^{-4}	8.72×10^{-4}	5.74×10^{-2}	5.74×10^{-2}

patients (first-tier sample) is the same as that from which the newborn (second-tier) sample is chosen. This is critical, since the values of r and p affect the estimation in two-tier screening surveys. For CF, r is known to vary widely between populations. For example, Lemna et al. (1990) noted that the proportion of CF alleles carrying the ΔF508 mutation is considerably higher (>75%) in northern European Caucasians than in Italy (57%), Portugal (53%), and Spain (51%). Within-country geographic variation in populations of the same racial background has also been noticed in the preliminary survey conducted in Brazil (S. Raskin, unpublished data). Therefore, unless appropriate caution is exercised in documenting the ethnic background of patients, one may easily derive incorrect estimates of the incidences of disease and gene carriers by using the above theory. Nevertheless, our observations that 50–100 patients per population and a 20-fold-larger sample of newborns would be adequate for diseases such as CF and MCAD should be of considerable significance in understanding the geographic distribution of recessive diseases. Such information should also provide insight into the origin and maintenance of recessive deleterious genes, and information thus collected should help to explain how such diseases have accumulated high frequencies in some regions of the world.

As mentioned earlier, this theory is also applicable to autosomal dominant diseases. In such cases, the disease frequency is given by

$$d^2 + 2d(1-d) = d(2-d) \approx 2d ,$$

when the disease is rare. This is almost equivalent to the frequency of disease gene bearers, $2d(1-d)$, for a recessive disease. Therefore, in principle the sample-size evaluations presented in table 2 should apply to the evaluation of the incidences of autosomal dominant diseases. However, all sample sizes reported in table 2 must be doubled, since, under the assumption that most affected individuals for a dominant disorder are heterozygous, and that, in contrast to the situation for recessive diseases, the mutation occurs on the disease allele(s)–bearing chromosome, we can count only one chromosome per individual for mutation screening, which results in twice the number of individuals needed. Nevertheless, computations presented in table 2 and figure 1*B* and *D* indicate that a mutation-screening program with 50–100 patients and 2,000 newborns should yield estimates of (dominant) disease incidence with a coefficient of variation around .25 or with 95% confidence intervals whose bounds will not differ by a factor of 5.

Acknowledgments

This work was supported by U.S. Public Health Service research grants GM41399 and GM45861 from the National Institutes of Health. Comments from Drs. W. J. Schull, E. Boerwinkle, and C. L. Hanis are greatly appreciated. We are thankful to Dr. Yixi Zhong for his programming help in preparing the tables and the figure.

References

Allan JL, Phelan PD (1985) Incidence of cystic fibrosis in ethnic Italians and Greeks and in Australians of predominantly British origin. Acta Pediatr Scand 74:286–289

Bennett MJ, Worthy E, Pollitt RJ (1987) The incidence and presentation of dicarboxlic aciduria. J Inherited Metab Dis 10:241–242

Boat TF, Welsh MJ, Beaudet AL (1989) Cystic fibrosis. In: Scriver AL, Beaudet AL, Sly WS, Valle D (eds) The metabolic basis of inherited disease. McGraw-Hill, New York, pp 2649–2680

Braekeller MD, Daigneault J (1992) Spatial distribution of DF508 mutation in cystic fibrosis: a review. Hum Biol 64:167–174

Chakraborty R, Srinivasan MR, Daiger SP (1993) Evaluation of standard error and confidence interval of estimated multilocus genotype probabilities, and their implications in DNA forensics. Am J Hum Genet 52:60–70

Cystic Fibrosis Genetic Analysis Consortium (1990) Worldwide survey of the ΔF508 mutation—report from the Cystic Fibrosis Genetic Analysis Consortium. Am J Hum Genet 47:354–359

Ferrie RM, Schwarz MJ, Robertson NH, Vaudin S, Super M, Malone G, Little S (1992) Development, multiplexing, and application of ARMS tests for common mutations in the CFTR gene. Am J Hum Genet 51:251–262

Kendall M (1947) The advanced theory of statistics. Vol 1. Charles Griffin, London

Kerem BS, Rommens JM, Buchanan DM, Cox TK, Chakravarti A, Buchwald M, Tsui L-C (1989) Identification of the cystic fibrosis gene: genetic analysis. Science 245:1073–1080

Kolvraa S, Gregersen N, Blakemore A, Scheiderman K, Winter V, Andersen B, Curtis D, et al (1991) The most common mutation causing medium-chain acyl-CoA dehydrogenase (MCAD) deficiency is strongly associated with a particular haplotype in the region of the gene. Hum Genet 87:425–429

Lemna WK, Feldman GL, Kerem BS, Fernbach SD, Zevcovich EP, O'Brien WE, Riordan JR, et al (1990) Mutation analysis for heterozygote detection and the prenatal diagnosis of cystic fibrosis. N Engl J Med 322:291–296

McKusick VA (1992) Mendelian inheritance in man, 10th ed. Johns Hopkins University Press, Baltimore

Newton CR, Graham A, Heptinstall LE, Powell SJ, Summers C, Kalsheker N, Smith JC, et al (1989) Analysis of any point mutation in DNA: the amplification refractory mutation system (ARMS). Nucleic Acids Res 17:2503–2516

Old JM, Varawalla NY, Weatherhall DJ (1990) Rapid detection and prenatal diagnosis of β-thalassemia: studies in Indian and Cypriot populations in the UK. Lancet 336:834–837

Rao CR (1973) Linear statistical inference and its applications. John Wiley, New York

Romeo G, Devoto M (1990) Population analysis of the major mutation in cystic fibrosis. Hum Genet 84:1–5

Sommer SS, Cassady JD, Sobell JL, Bottema CDK (1989) A novel method for detecting point mutations or polymorphisms and its application to population screening for carriers of phenylketonuria. Mayo Clin Proc 64:1361–1372

Wenham PR, Newton CR, Price WH (1991) Analysis of apoliprotein E genotypes by the amplification refractory mutation system. Clin Chem 37:241–244

Wu DY, Ugozzoli L, Pal BK, Wallace RB (1989) Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia. Proc Natl Acad Sci USA 86:2757–2760

Yokota I, Coates PM, Hale DE, Rinaldo P, Tanaka K (1991) Molecular survey of a prevalent mutation, ^{985}A-to-G transition, and identification of five infrequent mutations in the medium-chain-Acyl-CoA dehydrogenase (MCAD) gene in 55 patients with MCAD deficiency. Am J Hum Genet 49:1280–1291

Yokota I, Indo Y, Coates PM, Tanaka K (1990) Molecular basis of medium chain acyl-coenzyme A dehydrogenase deficiency: an A to G transition at position 985 that causes a lysine-304 to glutamate substitution in the mature protein is the single prevalent mutation. J Clin Invest 86:1000–1003