Am J Hum Genet 26:304–317, 1974

A Maximum-Likelihood Method for Estimating the Disease Predisposition of Heterozygotes

MICHAEL SWIFT,[1] JOSHUA COHEN,[2] AND ROGER PINKHAM[3]

INTRODUCTION

There are many ways to analyze genetic predisposition to common diseases such as cancer or diabetes. In most instances clinical disease results from the interaction of several factors, and it is difficult to isolate the effect of any single disease-predisposing gene. There are exceptions: rather uncommon single-gene syndromes clearly associated with a striking prevalence of cancer or diabetes in individuals who have the syndrome [1; 2, pp. 151–194].

One approach to detecting the disease-predisposing effect of single genes which are important in the general population is based on the hypothesis that a gene which produces a small increment in specific disease risk in the heterozygote may be associated with a recognizable syndrome in homozygotes [3, 4, 5]. There are several hundred known human autosomal recessive syndromes [6], and for most of them little is known of the disease tendencies of heterozygous carriers. Estimates of gene frequency suggest that for each of these recessive syndromes the heterozygote frequency is likely to be between .001 and .04. For any gene with a heterozygote frequency in this range, it is important to know whether the carriers of that gene are predisposed to any specific serious common disease. If we are interested in diabetes, for example, it seems sensible to examine the heterozygous carriers of genes causing recessive syndromes associated with carbohydrate intolerance [2, pp. 151–194]. It is unlikely that any single gene of this category produces an overwhelming tendency to diabetes in heterozygotes—that would have been noticed on casual inspection of pedigrees. A less obvious effect may still be important if diabetes is determined polygenically [2, pp. 151–194]: genes identified in homozygotes by the recessive syndromes they cause may comprise, in heterozygotes, a substantial proportion of the polygenic system.

Received July 2, 1973; revised November 12, 1973.

An earlier version of this report was read at the Fourth International Congress of Human Genetics, Paris, September 1971.

This research was supported by NIH grants CA 14235, GMO 1668, and HD 03110.

[1] Department of Medicine, Genetics Curriculum and the Biological Sciences Research Center, University of North Carolina, Chapel Hill, North Carolina 27514.

[2] Medical Scientist Trainee, New York University Medical Center, New York 10016.

[3] Department of Mathematics, Stevens Institutue of Technology, Hoboken, New Jersey 07030.

A direct test of this hypothesis is impossible at present, since, for most of the genes with which we are concerned, there is no reliable way to identify heterozygous carriers in the general population. For any gene associated with a recognizable autosomal recessive syndrome, we can, however, study the prevalence of diabetes (or any other common illness) in a group of obligatory heterozygotes, the parents or offspring of persons having the syndrome. This approach was used [7] to study Penrose's conjecture [3] that the phenylketonuric heterozygote is predisposed to mental illness and to search for an increased prevalence of specific chronic diseases in cystic fibrosis heterozygotes [5]. The sensitivity of such a comparison between a group of obligatory heterozygotes and a control group is determined by the size and composition of the experimental group. For example, in one study three cases of diabetes mellitus were found among 132 parents, aged 20–56, of cystic fibrosis probands [5]. In this age range, diabetes occurs in the United States population at a frequency below .01 [8]. If cystic fibrosis heterozygotes have two or three times the normal risk of developing diabetes, the most likely finding is the three cases of diabetes actually observed. But the difference between the observed number of diabetics in the experimental group (three) and the number expected in a sample from the general population (about one) is not significant. A much larger sample, or one composed of much older individuals, would be necessary to demonstrate a two- or threefold increase in risk of diabetes associated with heterozygosity for the cystic fibrosis gene.

To test the hypothesis that heterozygous carriers of the gene for the recessive syndrome Fanconi's anemia (FA) were unusually likely to die from a malignant neoplasm, death certificates were collected for all close relatives of eight FA probands (extending to first cousins once removed) who died after 1930 in the United States or Canada [4]. The strategy of including all relatives with a substantial probability of being heterozygous for the FA gene provided, compared to the available group of obligatory heterozygotes, a larger sample with a wider distribution of ages. A malignant neoplasm was found to have been the underlying cause of death in 27 out of 102 deaths, a finding which was compared to the 17.4 such deaths expected in a sample of that size for the general population.

The standard statistical techniques (χ^2, Poisson, binomial, etc.) for comparing the observed and expected number in this type of study ignore relevant information about the sample: the prior probability (by relationship to the proband) that a given relative is heterozygous for the gene in question. The assignments of probability of heterozygosity to relatives of probands are given in figure 1 for the cases in which the parents of the probands were either first cousins or totally unrelated; assignments for other types of matings are easy to deduce. In addition, the standard statistical techniques assess only the probability of observing a disparity (between the observed and control data) as large as the one in hand, while a statistic such as the relative risk better expresses the magnitude of disease predisposition for the heterozygous carrier.

The maximum-likelihood method which is presented here for estimating the relative risk to heterozygotes takes account of the probabilities of heterozygosity

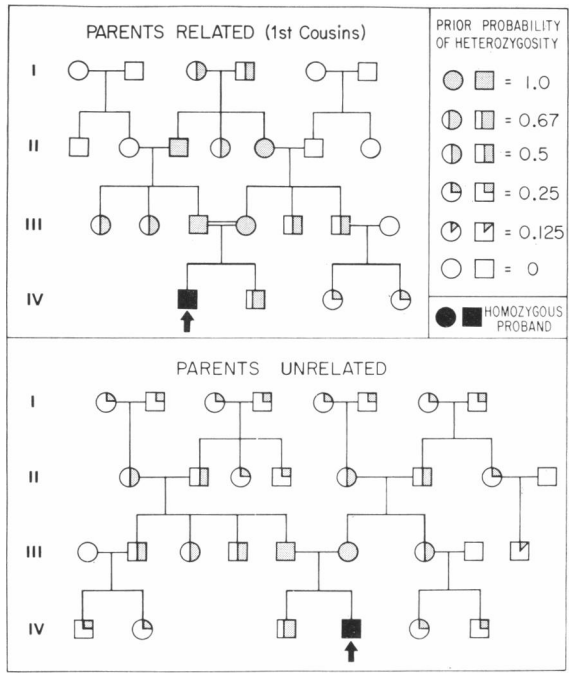

FIG. 1.—The prior probability that a relative of a proband with a rare autosomal recessive syndrome will be heterozygous for the gene is proportional to the gray-shaded portion of each circle or square. These probabilities of heterozygosity are based solely on the relationship to the homozygous proband and disregard the small probability of carrying the gene as a random member of the general population.

associated with various relationships to the proband and compares the observed incidence of disease to sex- and age-specific control data.

ANALYTICAL FORMULATION

In the general population the probability of a "test disorder" such as malignant neoplasm or diabetes is usually age-, race-, and sex-dependent, and we denote by R_j the probability of the test disorder for a particular age, race, and sex category. For a heterozygous carrier of a gene associated with an autosomal recessive syndrome, the probability of the test disorder is the product θR_j, where θ is the relative risk.

A relative of a homozygous proband will be of kind i (parent, sibling, uncle, etc.) and have probability P_i of being a heterozygous carrier (fig. 1). Then $P_i\theta R_j$ is the probability that a heterozygous carrier in age-race-sex class j who is a relative of kind i has the test disorder.

The probability of not being heterozygous but nonetheless suffering from the disorder in question is $(1 - P_i)R_j$. Thus finally

$$P_i\theta R_j + (1 - P_i)R_j \equiv \omega_{ij}$$

is the total probability of suffering from the disorder in question (malignant neoplasm, diabetes, etc.) for a relative of kind i in class j.

Note that the probability P_i is determined by theory from the family structure; R_j, on the other hand, is an empirical fact determinable in principle by census.

If the total number of relatives of kind i and class j is partitioned into x_{ij} who have the test disorder and y_{ij} who do not, then the probability of observing this event is given by the binomial distribution

$$\binom{x_{ij}+y_{ij}}{x_{ij}} \omega_{ij}^{x_{ij}} (1-\omega_{ij})^{y_{ij}}.$$

The probability for occurrences x_{ij} and nonoccurrences y_{ij} over all kinds and classes is supplied by the product

$$\prod_{i,j} \binom{x_{ij}+y_{ij}}{x_{ij}} \omega_{ij}^{x_{ij}} (1-\omega_{ij})^{y_{ij}}.$$

Now this product depends on θ only through the ω_{ij} and not the binomial coefficients. Thus it is convenient to omit these factors and define $L(\theta)$ by

$$L(\theta) = \prod_{i,j} \omega_{ij}^{x_{ij}} (1-\omega_{ij})^{y_{ij}}.$$

We refer to $L(\theta)$ as the likelihood function for the observations x_{ij}, y_{ij} (for all i and j).

MAXIMUM-LIKELIHOOD ESTIMATION

To estimate θ one may employ the method of maximum likelihood [9, 10], determining that value $\hat{\theta}$ of θ which maximizes $L(\theta)$. This may be done analytically but is often done empirically by making a plot of $\ln L(\theta)$ against θ and noting that value of θ which yields the maximum. Such estimates of θ are known to have many desirable properties [11, 12]. One property of particular value is that for large $\Sigma(x_{ij}+y_{ij})$, $\hat{\theta}$ can be shown to have a certain limiting normal distribution, and this allows one to make significance tests and confidence statements about θ [11].

In fact $\hat{\theta}$ is approximately normally distributed about θ with variance the reciprocal of

$$-E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right] = \sum_{i,j} \frac{(x_{ij}+y_{ij})P_i^2 R_j^2}{\omega_{ij}(1-\omega_{ij})},$$

where E denotes expected value. Thus the probability that

$$\theta - z\sigma \leqslant \hat{\theta} \leqslant \theta + z\sigma$$

is given by $\Phi(z)$, where

$$\sigma^2 = 1/-E\left[\frac{\partial^2 \ln L}{\partial \theta^2}\right],$$

and

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-z}^{z} e^{-\frac{1}{2}x^2} dx.$$

Examples of the use of this technique to find confidence intervals may be found in Kendall and Stuart [12]. We will demonstrate our use of this method in a subsequent paper and give computer simulations checking the quality of the approximations.

PRACTICAL APPLICATIONS

The relative risk that an FA heterozygote will die from a malignant neoplasm can be estimated using this maximum-likelihood method. In table 1 the 54 deaths of male FA relatives and 48 deaths of female FA relatives (from [4]) are partitioned by probability of heterozygosity and by underlying cause of death (malignant neoplasm or other cause). Two sets of control data, the first taken from the 1950 United States mortality data [13] and the second from the 1962 United States statistics, are also given in table 1. Each control datum, R_j, is the age-specific proportion of deaths due to malignant neoplasm among all deaths in the United States white male or female population for 1950 or 1962. The observed number of death certificates of FA relatives of kind i and age class j with a malignant neoplasm as the cause of death is x_{ij}, and the number of deaths from all other causes of death in each subgroup is given by y_{ij}. The procedure for abstracting the underlying cause of death from each death certificate of an FA relative was the same as that used in compiling the control standard mortality statistics.

The ln $L(\theta)$ for the male FA relatives and the 1950 control data is plotted in figure 2. Because ln $L(\theta)$ is maximum at $\theta = 3.2$, we estimate that a male FA heterozygote has 3.2 times the normal risk of dying from a malignant neoplasm (table 2). All risk estimates derived from the FA family data are summarized in table 2. Using the 1962 mortality statistics as control data and the same set of observed x_{ij} and y_{ij}, we estimated the relative risk of malignant neoplasm for males to be 2.6 (compared to 3.2 when the 1950 control data were used) and for females, 1.1 (1.5 when the 1950 control data were used). (But see the note added in proof, p. 316.)

Swift et al. [14] also analyzed the illness and mortality data of the FA families for the prevalence of diabetes mellitus. While conventional mortality statistics [13] (by single underlying cause) are a reliable and consistent measure of the prevalence of malignant neoplasms among the dying [15], for diabetes, surveys that tabulate all the causes of death and associated conditions [16] reflect the prevalence of this diagnosis at death much more accurately [17]. With both the FA death certificates [14] and the control data (the survey of all causes of death on United States death certificates in 1955 [16]) analyzed in this way (table 3), the risk that FA heterozygotes die with diabetes was estimated to be 2.1 times normal for the males and 7.3 for the females.

For the living FA relatives, control data from three different population surveys were used in three separate sets of estimates of risk of diabetes for male and female

TABLE 1

DEATHS FROM MALIGNANT NEOPLASMS AMONG RELATIVES OF FA PROBANDS

AGE GROUP	CONTROL RATE (R_j) U.S. 1950	CONTROL RATE (R_j) U.S. 1962	PROBABILITY OF HETEROZYGOSITY (P_i) .125 x_{ij}	.125 y_{ij}	.25 x_{ij}	.25 y_{ij}	.5 x_{ij}	.5 y_{ij}	.667 x_{ij}	.667 y_{ij}	1.0 x_{ij}	1.0 y_{ij}
						Male						
0–4	.0163	.0175	...	...	...	1	...	...	...	...	...	...
5–9	.1237	.1793	...	...	1	...	...	...	...	...	...	...
20–24	.0603	.0726	...	1	...	...	...	1	...	...	...	...
30–34	.1026	.1228	...	...	...	...	...	...	...	1	...	...
35–39	.1115	.1289	...	1	...	...	...	...	...	...	...	...
40–44	.1204	.1485	1	3	...	1	...	...	...	...	...	...
45–49	.1427	.1685	...	...	...	1	...	...	...	...	...	...
50–54	.1605	.1872	1	...	...	1	...	1	...	...	...	...
55–59	.1743	.2053	...	...	...	3	2	...	...	...	...	...
60–64	.1804	.2085	...	...	4	2	...	5	...	...	1	...
65–69	.1686	.1974	...	...	1	1	...	3	...	...	...	...
70–74	.1599	.1779	...	1	2	2	...	2	...	...	...	...
75–79	.1393	.1507	...	1	1	3	1	1	...	...	...	...
80–84	.1177	.1224	...	...	...	1	...	...	...	...	...	...
85+	.0784	.0843			...	2						
						Female						
0–4	.0182	.0200	...	...	...	2	...	2	...	...	...	...
15–19	.1055	.1293	...	1	...	1	...	...	...	...	...	...
25–29	.1521	.1650	...	...	...	...	...	1	...	...	...	...
35–39	.2870	.3101	...	2	...	...	...	...	...	...	...	...
40–44	.3372	.3580	...	1	...	1	...	...	...	...	...	...
50–54	.3362	.3799	2	1	1	...	...	...	...	...	...	...
55–59	.3091	.3366	1	...	1	4	1	1	...	...	...	1
60–64	.2588	.2895	...	1	1	...	...	...	...	...	...	...
65–69	.2119	.2390	...	...	2	5	1	1	...	...	...	...
70–74	.1719	.1821	...	...	1	1	1	...	...	...	...	...
75–79	.1357	.1391	...	1	...	1	...	3	...	...	...	...
80–84	.1043	.1030	...	...	...	4	...	...	...	...	...	...
85+	.0686	.0652	...	...	...	1	...	...	...	...	...	...

NOTE.—For each relative of an FA proband in the families studied [4], P_i is determined solely by his or her relationship to that proband (see fig. 1). Each R_j is the age-specific proportion of deaths due to malignant neoplasm among white males or females taken from the United States mortality statistics for 1950 or 1962 [13]. As discussed in the text, a single set of R_j is used in each estimation of θ. Each x_{ij} is the number of deaths from malignant neoplasm observed for relatives of FA probands of probability group i and age class j, and y_{ij} is the number of observed deaths due to all other causes for relatives of the same sex in the same group and class.

FA heterozygotes. The control and observed data are listed in tables 4 and 5, and the estimates of θ for each set of control data are in table 2. The control and experimental data were not obtained by identical survey techniques, since the FA relatives were asked to list their hospitalizations and illnesses, while the controls were asked directly whether they had diabetes. Thus the values in table 2 for the relative risk of diabetes for living FA heterozygotes may be underestimates.

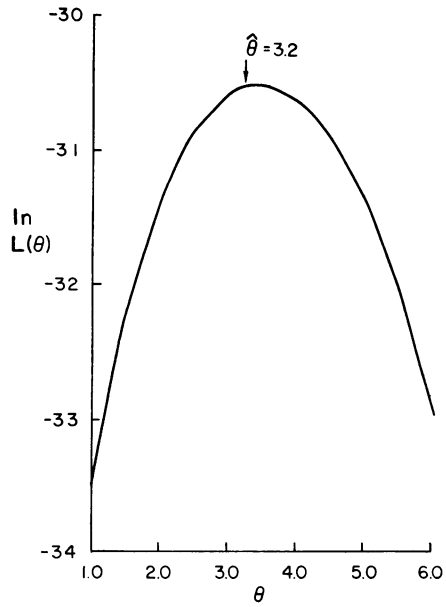

FIG. 2.—The relative risk of dying from a malignant neoplasm for a male FA heterozygote was estimated by plotting ln $L(\theta)$ against θ, using the likelihood function described in the text and the observed and control data (1950 mortality statistics) in table 1.

CONTROL DATA

Since the observed data are compared to values established by census or survey from the general population, the most serious question is whether the experimental families may be regarded as a random sample from the general population. While the FA families were chosen only because they contained one or more probands with Fanconi's anemia, it might be that this particular syndrome occurs mainly in a population subgroup in which malignant neoplasms are unusually frequent, independent of the FA gene. If this were true, the increased risk of dying from malignant neoplasm might not be associated with the FA gene. Studies of medical-genetic associations are often questioned because of doubts about the stratification of the sample.

There is no absolute answer to such doubts. Obvious stratification might be detected by examining the ethnic origins, social class, occupations, or geographical distribution of the families in the sample. Previous studies of medical-genetic associations have used special groups as additional controls for hidden stratification in the experimental group. "Sib" controls [19] would not be appropriate here, but "spouse" controls would. The spouses of sibs of affected probands, and of the grandparents' sibs, aunts, uncles, and cousins, are not part of the experimental sample. Since they presumably came from the same population strata as the family members within the sample, they constitute a suitable control group. The value of this control is limited by its small size and restricted age distribution.

TABLE 2

ESTIMATES OF THE RELATIVE RISK OF MALIGNANT NEOPLASM OR OF DIABETES FOR FA HETEROZYGOTES

	HETEROZYGOTE GROUP			
	Sex	Age range	CONTROL DATA AND REFERENCE	ESTIMATE OF θ
I. Deaths from malignant neoplasms ..	Male	All	Mortality by underlying cause, U.S. 1950 [13]	3.2
	Male	All	Mortality by underlying cause, U.S. 1962 [13]	2.6
	Male	>30	Mortality by underlying cause, U.S. 1950 [13]	3.2
	Male	>55	Mortality by underlying cause, U.S. 1950 [13]	3.4
	Male	All	Mortality by underlying cause, U.S. 1950 [13]	3.1*
	Female	All	Mortality by underlying cause, U.S. 1950 [13]	1.5
	Female	All	Mortality by underlying cause, U.S. 1962 [13]	1.1
	Both	All	Mortality by underlying cause, U.S. 1950 [13]	2.1
II. Deaths with diabetes ...	Male	All	Survey of multiple causes of death, U.S. 1955 [16]	2.1
	Male	>35	Survey of multiple causes of death, U.S. 1955 [16]	2.4
	Female	All	Survey of multiple causes of death, U.S. 1955 [16]	7.3
	Female	>35	Survey of multiple causes of death, U.S. 1955 [16]	7.5
	Female	All	Survey of multiple causes of death, U.S. 1955 [16]	6.8*
III. Diabetes in living individuals	Male	All	U.S. National Health Interview Study 1964–1965 [8]	0.9
	Female	All	U.S. National Health Interview Study 1964–1965 [8]	5.3
	Female	>25	U.S. National Health Interview Study 1964–1965 [8]	5.4
	Female	>45	U.S. National Health Interview Study 1964–1965 [8]	5.9
	Female	All	U.S. National Health Interview Study 1964–1965 [8]	5.2*
	Female	All	Prince Edward Island, Canada 1966 [20]	5.9
	Female	All	Sudbury, Massachusetts, 1965 [18]	9.7

NOTE.—All estimates of θ were obtained by the maximum-likelihood method described in the text from the data tabulated in tables 1 and 3–5. Estimates of θ are given for particular groupings of heterozygotes (by sex, age, or probability of heterozygosity) and for different sets of population control statistics.

* Estimate of θ when data were omitted for relatives whose probability of heterozygosity was less than .25.

A set of families "matched" to the experimental families for various stratification parameters would not provide reliable control data, since such control families would not provide data of accuracy and completeness comparable to those obtained from the experimental families. It has been shown, in fact, that malignant neoplasms in families "matched" to those with breast cancer probands were found at a prevalence lower than that of the general population [21], strongly suggesting that such controls were inaccurate or incomplete in their reporting.

It is clear from the nature of the likelihood function and from the examples given above that, for a given set of observed data, the estimate of relative risk varies if the control data are changed. Often it is necessary to choose among sets of control data, since several may appear to be equally valid representations of the state of affairs in the control population. The interpretation of differences in estimates of θ when different sets of control data are used depends on their magnitude. Small

TABLE 3

DEATHS WITH DIABETES IN RELATIVES OF FA PROBANDS

AGE GROUP	CONTROL RATE (R_j)	PROBABILITY OF HETEROZYGOSITY (P_i)									
		.125		.25		.5		.667		1.0	
		x_{ij}	y_{ij}	x_{ij}	y_{ij}	x_{ij}	y_{ij}	x_{ij}	y_{ij}	x_{ij}	y_{ij}
		Male									
0–4	.0007	...	...	...	1	...	...	...	...	...	...
5–14	.0078	...	...	...	1	...	...	...	...	...	...
15–24	.0071	...	1	...	...	...	1	...	...	...	...
25–34	.0185	...	...	...	...	...	...	...	1	...	...
35–44	.0210	...	5	...	1	...	...	...	...	...	...
45–54	.0240	...	1	...	2	...	1	...	...	...	...
55–64	.0371	...	...	1	8	...	7	...	...	...	1
65–74	.0411	...	1	1	5	...	5	...	...	...	...
75–84	.0333	...	1	1	4	...	2	...	...	...	...
85+	.0182	...	...	...	2	...	...	...	...	...	...
		Female									
0–4	.0006	...	...	...	2	...	2	...	...	...	...
5–14	.0142	...	...	...	...	...	...	...	...	...	...
15–24	.0242	...	1	...	1	...	...	...	...	...	...
25–34	.0294	...	...	...	...	...	1	...	...	...	...
35–44	.0238	1	2	...	1	...	...	...	...	...	...
45–54	.0440	...	3	...	1	...	...	...	...	...	...
55–64	.0883	1	1	2	4	1	1	...	...	...	1
65–74	.0949	...	...	4	6	...	2	...	...	...	...
75–84	.0564	...	1	2	3	...	3	...	...	...	...
85+	.0234	...	...	...	1	...	...	...	...	...	...

NOTE.—P_i is described in the note to table 1. Each R_j is the age-specific proportion of deaths of white males or females for whom diabetes was listed as one of the causes of death or associated conditions on the death certificate, as reported in the 1955 survey of multiple causes of deaths on death certificates [16]. Each x_{ij} is the number of deaths of relatives of FA probands of probability group i and age class j for whom diabetes was listed on the death certificate, and y_{ij} is the number of death certificates remaining among relatives of the same sex in that group and class.

differences in relative risk are biologically unimportant and may be disregarded, while larger discrepancies between estimates must be investigated.

There is, for example, a relatively large difference between the estimate of risk of diabetes for living female FA heterozygotes based on the United States Health Survey [8] data, 5.3, and the estimate using the Sudbury, Massachusetts [18] data, 9.7. The Sudbury study was, however, noteworthy for finding a much lower prevalence of diabetes in adult females than almost all other surveys of comparable populations. Comparisons such as this are often available to check on the validity of control data and of risk estimates derived from them.

HETEROGENEITY OF RELATIVE RISK

In the general population the absolute risk of developing a test disorder such as cancer or diabetes varies enormously with age, sex, or other population parameters.

TABLE 4

DIABETES AMONG LIVING RELATIVES OF FA PROBANDS

AGES	PROBABILITY OF HETEROZYGOSITY (P_i)									
	.125		.25		.5		.667		1.0	
	x_{ij}	y_{ij}	x_{ij}	y_{ij}	x_{ij}	y_{ij}	x_{ij}	y_{ij}	x_{ij}	y_{ij}
Male										
0–4	...	7	...	3	...	...	...	1	...	...
5–9	...	4	...	2	...	...	...	2	...	...
10–14	...	11	...	4	...	...	...	3	...	...
15–19	...	3	...	6	...	...	...	2	...	...
20–24	...	4	...	5	...	...	...	...	...	...
25–29	...	2	...	8	...	...	...	...	...	...
30–34	...	3	...	1	...	2	...	...	...	...
35–39	...	4	...	2	...	3	...	...	...	...
40–44	1	10	...	2	...	2	...	...	...	1
45–49	1	3	...	2	...	3	...	...	...	4
50–54	...	7	...	3	...	4	...	...	...	1
55–59	...	9	...	1	...	3	...	...	...	1
60–64	...	5	...	2	...	1	...	...	...	1
65–69	...	3	...	3	...	3	...	...	...	...
70–74	...	1	...	3	...	...	...	...	...	...
75–79	...	...	...	2	...	2	...	...	...	...
80+	...	...	2	1	...	...	...	...	...	...
Female										
0–4	...	2	...	4	...	1	...	...	...	...
5–9	...	7	...	4	...	...	...	1	...	...
10–14	...	4	...	7	...	...	...	2	...	...
15–19	...	5	...	4	...	1	...	2	...	1
20–24	...	5	...	12	...	3	...	...	...	...
25–29	...	4	...	4	...	1	...	...	...	...
30–34	...	9	...	...	...	3	...	...	...	...
35–39	...	9	...	5	...	1	...	...	...	...
40–44	...	11	...	1	...	2	...	...	...	4
45–49	...	16	...	2	...	1	...	...	1	1
50–54	...	6	...	1	...	6	...	...	...	...
55–59	...	5	...	1	...	5	...	...	...	1
60–64	1	5	...	2	...	3	...	...	...	1
65–69	1	2	...	4	2	2	...	...	...	...
70–74	...	...	1	3	1	2	...	...	...	...
75–79	...	...	1	2	...	3	...	...	...	...
80+	...	...	...	1	...	...	...	...	...	...

NOTE.—The interpretation of P_i is given in the note to table 1. Each x_{ij} is the number of diagnosed cases of diabetes mellitus among relatives of FA probands of probability group i and age class j, and y_{ij} is the number of other relatives of the same sex in the same group and class. The control data, R_j, are given in table 5.

Such variation in risk for the overall population is incorporated into the maximum-likelihood estimation of relative risk described and demonstrated above. Furthermore, for a given gene the relative risk of the test disorder may also vary as these or other parameters are varied. Such variation can be seen in the data analyzed in Practical Applications and summarized in table 2. For FA heterozygotes there is

TABLE 5

PREVALENCE OF DIABETES IN THREE NORTH AMERICAN POPULATIONS

POPULATION, REFERENCE, AND AGE GROUP	PROPORTION OF DIABETICS (R_j)	
	Male	Female
U.S. Health Survey, 1964–1965 [8]:		
<25	.0012	.0013
25–44	.0062	.0062
45–54	.0154	.0200
55–64	.0320	.0414
65–74	.0471	.0606
75+	.0470	.0508
Prince Edward Island, Canada, 1966 [20]:		
0–9	.0005	.0003
10–19	.0013	.0015
20–29	.0027	.0020
30–39	.0032	.0050
40–49	.0050	.0075
50–59	.0194	.0248
60–69	.0262	.0460
70–79	.0381	.0604
80–89	.0433	.0580
90+	.0400	.0350
Sudbury, Massachusetts, 1965 [18]:		
15–24	.000	.000
25–34	.006	.003
35–44	.011	.000
45–54	.034	.007
55–64	.036	.023
65–74	.139	.033
75+	.053	.059

NOTE.—Each population survey was used separately in estimating the relative risk of diabetes for FA heterozygotes from the data in table 4. The estimates obtained are listed in table 2. Each R_j is the age- and sex-specific proportion of diagnosed diabetics discovered in that survey, grouped by age as they were originally reported.

a substantial difference between men and women in the estimates of relative risk for diabetes. The significance and biological meaning, if any, of this difference in risk estimates between the sexes are unclear.

While it is desirable in principle to examine relative risk estimates for male or female heterozygotes grouped by age, it is obvious that small sample sizes would likely lead to poor statistical precision. The proper procedure in designing a study would be to choose in advance groupings by age and sex that seem appropriate to the problem at hand and to insist that pertinent subsamples be of adequate size. For data that have already been collected, such as those presented in this paper, it is interesting to examine the effect of subdividing the sample by age category (table 2). As we have seen, the estimated relative risks for males and females, based on the reported data, differ widely. In contrast, if one estimates relative risks for restricted age groups within each sex, no great disparity in relative risk appears. If the sample size were large enough and if there were an important difference for

male or female heterozygotes in the relative risk of the test disorder at different ages, then this manner of analysis would reveal it. It may be possible to examine how well an estimate of θ fits a body of data by the likelihood ratio or other test of goodness of fit. We are presently studing the conditions under which such a test can be applied.

CONCLUSION

A numerical value for the relative risk of developing some test disorder usefully summarizes information about disease predisposition associated with a particular gene. The method of data collection described previously [4] and the mathematical analysis presented in this paper offer a practical way of estimating the disease-predisposing effects of each of the several hundred human genes which can be identified at present only by the rare syndrome which they produce in homozygotes.

The confidence limits for estimates of relative risk will be analyzed in subsequent papers of this series. Still, it is important to note that these risk estimates have only limited numerical precision. In general, we wish to know the magnitude of elevation in disease risk associated with a particular gene—whether it is two, three, four, or more times normal—and the significance of the observed elevation.

While the Fanconi's anemia study [4, 14] may serve to illustrate the usefulness of this method, there are deficiencies in it which should be avoided in the future. Except for syndromes of extreme rarity, it should be possible to study more than eight families, thus providing a larger sample. Tracing relatives with a probability of .125 of being heterozygous was difficult out of proportion to the value of the information obtained, and examples in table 2 show that omitting relatives at the .125 probability level changed the relative risk estimates very little. We do not plan to include relatives at this probability level in future studies. Also, spouse controls should be compiled along with blood relatives, even though the data on these spouses will be limited in accuracy, quantity, and scope.

Another limitation in estimating relative risk derives from using the "proportion of" deaths from the test disorder. For example, for women in the general population aged 35–60, about 0.33 of all deaths are due to a malignant neoplasm. No matter how many deaths from malignant neoplasm occurred in carriers of a particular gene, it is impossible to obtain (using these age-specific proportionate death rates) an estimate of relative risk greater than three for women of this age. A more accurate estimate of risk might be obtained by comparing observed and control incidence rates in the maximum-likelihood calculation. (See the note added in proof, p. 316.) Finally, the method demonstrated in this paper does not analyze the effect of competing risks.

For an uncommon gene randomly distributed in the population, the fraction of heterozygous gene carriers with a specific test disorder among all patients with that disorder is given approximately by the product of the heterozygote frequency and the heterozygote's relative risk of the disorder. For example, if the frequency of FA heterozygotes in the population is .003 [4] and the relative risk of malignant neoplasm for FA heterozygotes is about three, then we estimate that 1% of all

persons dying from cancer and leukemia would carry the FA gene. Similarly, if the relative risk of diabetes for female FA heterozygotes is six, then about 2% of all female diabetics carry the FA gene.

When specific biochemical tests become available to identify the heterozygous carriers of genes causing autosomal recessive syndromes, estimates of relative risk for each such gene can be based on direct measurements of the heterozygote frequency in the general population and in groups afflicted with the specific test disorder. For the time being it may be desirable to measure the disease-predisposing effects of these genes by the indirect methods presented here.

SUMMARY

Genes which produce autosomal recessive syndromes may have a significant effect in predisposing heterozygous carriers to serious common diseases such as cancer or diabetes. For any specific autosomal recessive syndrome, such disease predisposition may be detected, if it exists, by studying the major illnesses and causes of death in the families of homozygous probands. Each relative has, by virtue of his relationship to the proband, a prior probability of being heterozygous for the gene for the syndrome. We use these prior probabilities in a maximum-likelihood method to estimate from the compiled family data the risk that the heterozygote will develop or die from the common illness under investigation.

Genes which can be studied in this way may account for a substantial proportion of disease predisposition in the general population.

ACKNOWLEDGMENTS

We thank Derek Hudson, Bruce Hoadley, Brenda Edwards, and Douglas Gilmour for helpful discussions.

NOTE ADDED IN PROOF.—Since this article was submitted for publication, a computer program was developed to extract information about the *incidence* of death from specific causes from family data collected in the Fanconi's anemia and similar studies. As indicated in the text, estimations of risk based on incidence data may measure disease predisposition more reliably than those employing age-specific proportionate death rates.

We estimated the relative risk of dying from a malignant neoplasm, comparing the incidence in the FA families to 1960 control rates for the U.S. white population [22], to be 3.4 for male and 2.6 for female FA heterozygotes. With this more accurate method of estimating risk, there is a small difference in risk increment between the sexes.

REFERENCES

1. LYNCH HT: *Hereditary Factors in Carcinoma.* New York, Springer-Verlag, 1967
2. RIMOIN DL, SCHIMKE RN: *Genetic Disorders of the Endocrine Glands.* Saint Louis, Mosby, 1971
3. PENROSE LS: Inheritance of phenylpyruvic amentia (phenylketonuria). *Lancet* 2:192–194, 1935
4. SWIFT M: Fanconi's anaemia in the genetics of neoplasia. *Nature* (Lond) 230:370–373, 1971

5. HALLETT WY, KNUDSON AG, MASSEY FJ: Absence of detrimental effect of the carrier state for the cystic fibrosis gene. *Am Rev Respir Dis* 92:714–724, 1965
6. MCKUSICK VA: *Mendelian Inheritance in Man.* Baltimore, Johns Hopkins Univ. Press, 1971
7. PERRY TL, TISCHLER B, CHAPPLE JA: The incidence of mental illness in the relatives of individuals suffering from phenylketonuria or mongolism. *J Psychiatr Res* 4:51–57, 1966
8. U.S., DEPARTMENT OF HEALTH, EDUCATION AND WELFARE, NATIONAL CENTER FOR VITAL AND HEALTH STATISTICS: *Characteristics of Persons with Diabetes: United States, July 1964 to June 1965.* Public Health Service Publication no. 1000, ser. 10, no. 40, Washington, D.C., Government Printing Office, 1967
9. FISHER RA: *Statistical Methods for Research Workers.* New York, Hafner, 1958
10. FISHER RA: *Statistical Methods and Scientific Inference.* London, Oliver & Boyd, 1956
11. CRAMER H: *Mathematical Methods of Statistics.* Princeton, N.J., Princeton Univ. Press, 1951
12. KENDALL MG, STUART A: *The Advanced Theory of Statistics.* London, Griffin, 1961
13. U.S., DEPARTMENT OF HEALTH, EDUCATION AND WELFARE, NATIONAL OFFICE OF VITAL STATISTICS: *Mortality Statistics,* vol 2. Washington, D.C., Government Printing Office, 1950, 1962
14. SWIFT M, SHOLMAN L, GILMOUR D: Diabetes mellitus and the gene for Fanconi's anemia. *Science* 178:308–310, 1972
15. JABLON S, ANGEVINE DM, MATSUMOTO YS, ISHIDA M: On the significance of cause of death as recorded on death certificates in Hiroshima and Nagasaki, Japan. *Natl Cancer Inst Monogr* 19:445–465, 1966
16. U.S., DEPARTMENT OF HEALTH, EDUCATION AND WELFARE, NATIONAL CENTER FOR HEALTH STATISTICS: *Vital Statistics of the United States,* suppl.: *Mortality Data, Multiple Causes of Death.* Washington, D.C., Government Printing Office, 1955
17. OLSON FE, NORRIS FD, HAMMES LM, SHIPLEY PW: A study of multiple causes of death in California. *J Chronic Dis* 15:157–170, 1962
18. O'SULLIVAN JB, WILLIAMS RF, MCDONALD GW: The prevalence of diabetes mellitus and related variables—a population study in Sudbury, Massachusetts. *J Chronic Dis* 20:535–543, 1967
19. CLARKE CA, EDWARDS JW, HADDOCK DRW, HOWEL-EVANS AW, MCCONNELL RB, SHEPPARD, PM: ABO blood groups and secretor character in duodenal ulcer: population and sibship studies. *Br Med J* 2:725–31, 1956
20. SIMPSON NE: Diabetes in the families of diabetics. *Can Med Assoc J* 98:427–432, 1968
21. PENROSE LS, MACKENZIE HJ, KARN MN: A genetical study of human mammary cancer. *Br J Cancer* 2:168–176, 1948
22. U.S., DEPARTMENT OF HEALTH, EDUCATION AND WELFARE, NATIONAL CENTER FOR HEALTH STATISTICSS *United States Life Tables by Causes of Death: 1959–61.* Washington, D.C., Government Printing Office, 1968