J Chron Dis Vol. 35, pp. 539 to 551, 1982
Printed in Great Britain.

0021-9681/82/070539-13$03.00/0

PROBLEMS IN THE ASSESSMENT OF RELATIVE RISK OF CHRONIC DISEASE AMONG BIOLOGICAL RELATIVES OF AFFECTED INDIVIDUALS

KENNETH M. WEISS, RANAJIT CHAKRABORTY, PARTHA P. MAJUMDER
Center for Demographic and Population Genetics, University of Texas Health Science Center, Houston, TX 77025, U.S.A.

and

PETER E. SMOUSE
Department of Human Genetics, University of Michigan, Ann Arbor, Michigan, U.S.A.

(*Received in revised form* 25 *September* 1981)

Abstract—A question often asked in regard to a chronic disease is whether the risk to a biological relative of a case is elevated, and if so by how much the risk is altered. To answer this question, data may be collected directly with genetic objectives in mind by ascertaining populations of pedigrees. More often, the initial assessment of the question comes from family history data collected in an incidental manner in the course of a case–control or similar type of study. This paper discusses some limitations to the inferences which can be derived from such casual family history data. These include (i) poor statistical properties of standard relative risk measures, (ii) interpretational problems of observed relative risks when affected cases arise from genetic as well as nongenetic causes and when genes may not always be expressed in individuals in whom they are present, and (iii) confounding effects which may occur when a high risk allele alters the age of onset pattern of the disease. These problems result largely from a loss of design control over the degree of exposure of individuals ascertained and can lead to a small observed relative risk value even when genetic factors are important. Suggestions for handling such casual family history data are offered.

INTRODUCTION

A BASIC problem in contemporary epidemiology is the identification and evaluation of risk factors for chronic diseases such as cancers, cardiovascular disease, diabetes, etc. or perhaps for precursor states of these diseases. One often seeks environmental factors related to life style, geography, diet, occupation, and so on, and the search is accomplished through a case–control design. To make such a search complete, family data on the disease state of close relatives of cases and controls are (or should be) gathered, at least to ensure that genetic risk does not confound the evaluation of environmental variables. As distinct from detailed pedigree data gathered explicitly to detect familial aggregation, we will refer to family data collected in a general risk survey as 'casual' family history data.

When casual family history data of this kind are used, however, a fundamantal change in the analytic framework of the study occurs. Although gathered in retrospective terms, the family risk data are actually prospective in nature; they are studies of the exposure of the *relative* to the genes (or common family environment) of the proband. The cases and controls (viewed from this vantage point) become the *exposing*, not the exposed, factors. The duration of exposure is the age of the relative and the dose level is the degree of genetic relationship to the proband, that is, the probability of shared genes with the proband at any specific locus; because of Mendelian segregation of genes, only a fraction of the relatives are exposed. The outcome of such a study can be arrayed as in Table 1. Note that while the basic study on environmental risk may have been designed for a pre-determined number of cases and controls, there is *no* such pre-determination of the

TABLE 1. CONTINGENCY TABLE FORMAT FOR FAMILY HISTORY DATA

	Affected	Not affected	Total
Relative of:			
Case	n_{11}	n_{12}	N_1
Control	n_{21}	n_{22}	N_2
Total	M_1	M_2	T

numbers of observations in the family risk table, that is, of the numbers of relatives who are ascertained.

Risk may be quantified in various ways, but is usually expressed as either the Relative Risk, $RR = (n_{11}/N_1)/(n_{21}/N_2)$ or the Odds Ratio, $OR = (n_{11}/n_{12})/(n_{21}/n_{22})$. These two measures have different applications in case–control and cohort studies [1, 51], but to be general we will refer to them collectively as 'relative risks' unless otherwise specified. It is usually assumed that there is an underlying risk P_1 to the relatives of cases, estimated by n_{11}/N_1, and a different risk P_2 to relatives of controls, estimated by n_{21}/N_2, in the population from which the samples were drawn. The object of the analysis is (1) to *detect* a familial excess risk ($P_1 > P_2$), and (2) to assess the *strength of effect* of proband disease status on the risk to the relative, that is, the amount of difference between P_1 and P_2. Standard χ^2 tests can be used to determine whether RR or OR are significantly different from their no-association value of 1.0 under the null hypothesis of no familial excess risk [1–3, 51]. Estimating the difference in risk, however, is often not distinguished from the more subjective assessment of the importance of genetic factors in the etiology of the disease.

We probably have already discovered most of the clearly genetically determined disorders [4]. Those that remain will be difficult to disentagle using statistical data alone; their understanding will require a biochemical, physiological, or molecular characterization instead, but this depends on the initial detection of familial clusters so that high risk individuals may be identified for detailed study. It is important to search for familial risk whenever a risk factor survey is undertaken, especially because primary genetic lesions are more and more frequently implicated in degenerative diseases.

Several years ago, Lilienfeld noted as curious the fact that although there was persistent evidence for familial breast cancer, relative risks in studies of families of breast cancer patients vs controls always seemed to be small, about twofold [5]; similar small values have been observed for many chronic diseases, for example, lung cancer [6], colon cancer [7–10], and adult-onset diabetes mellitus [8]. Small relative risks like these can be interpreted to mean that genetics are not important in the causation or epidemiology of the diseases for which they were derived. However, the nature of the genetics of chronic diseases, and of relative risks as means to assess the genetics, are such that small familial relative risks are to be expected much of the time. In this paper we examine some of the reasons why this is true.

1. THE SIMPLEST CASE: ETIOLOGICALLY HOMOGENEOUS SINGLE-GENE TRAITS

The more complicated the etiological phenomena of a disease are, the more careful one must be in attempting to infer causal genetic elements. Even in the simplest case, when a disease is caused by an allele at a single genetic locus, and when there are no other loci, no age effects, and no environmental causal factors, care must be taken in interpreting casual family data. An examination of the relative risks produced by such a simple genetic circumstance will reveal some of the cautions which must be heeded in genetic epidemiology; in more complex situations there are additional problems, as will be shown later.

TABLE 2. PROBABILITY A RELATIVE (Y) OF A PROBAND (X) IS AFFECTED AS A FUNCTION OF DISEASE STATUS OF PROBAND AND DEGREE OF KINSHIP, FOR AUTOSOMAL DOMINANT AND RECESSIVE GENES

Type of relative	Probability relative is affected	
	Autosomal dominant	Autosomal recessive
Parent X–Offspring Y		
Pr(Y is Af.|X is Af.)	$(1 + p - p^2)/(2 - p)$	p
Pr(Y is Af.|X is Nor.)	p	$p^2/(1 + p)$
Sib X–Sib Y		
Pr(Y is Af.|X is Af.)	$(4 + 5p - 6p^2 + p^3)/(4(2 - p))$	$(1 + p)^2/4$
Pr(Y is Af.|X is Nor.)	$p(4 - p)/4$	$p^2(3 + p)/(4(1 + p))$
Uncle X/Aunt X–Nephew Y/Niece Y		
Pr(Y is Af.|X is Af.)	$\frac{1}{2}((1 - q^2) + (p/(1 - q^2)) + (pq/(1 + q)))$	$p(1 + p)/2$
Pr(Y is Af.|X is Nor.)	$p(2 + q)/2$	$p^2q/(1 - p^2)$
First Cousin X–First Cousin Y		
Pr(Y is Af.|X is Af.)	$(p/4(1 - q^2))(4p^3 + q(15p^2 + 14pq + 1))$	$p(q + 4p)/4$
Pr(Y is Af.|X is Nor.)	$p(4 + 3q)/4$	$(p^2q/(1 - p^2))(1 + 3p/4)$

Based on Li and Sacks' [14] ITO-matrix method. p is frequency of disease-related gene, say D, in the population; $q = 1 - p$ is the frequency of the alternative gene (say d) at that locus. To be affected, if disease is dominant, individual must have genotype either homozygous DD or heterozygous Dd; if disease is recessive, individual must have genotype DD.

The first point to be borne in mind in using casual family data is that different modes of inheritance will appear differently in specific sets of relatives. For example, a recessive disorder will generally not be manifest in parent–offspring data unless the disorder is very common or sample sizes are very large, but rather will be found mostly in sibs descended from unaffected parents. A sex-linked recessive disease will usually be observed to cluster in grandfather-grandson pairs. Because the fraction of shared genes declines with decreasing degree of kinship, there will be a trend in the expected excess relative risk with different degrees of relationship. Such a trend itself may be of great help in sorting out genetic from environmental familial factors, as the latter may not produce such a trend. Although χ^2 tests on the contingency table can be used to detect risk excess, the amount of excess, and hence the power to detect it with a given sample size, will depend on the type of relative involved.

The probability that a relative of a proband is affected depends, even in the simplest purely genetic situation, on several factors: the type of relationship, the affection status of the proband, and the disease frequency in the population. In general, for a single gene trait, the population frequency of an allele producing the disorder can be computed from the disease frequency in the population, π, and a knowledge of which genotypes are affected [12]; while such knowledge is not available in studies of diseases of unknown etiology, it can be applied here to illustrate the effect of genetics on relative risks. For example, if there are two alleles, A and a, at a locus and if the a allele produces a disorder only in the aa homozygote, then the disease is recessive and the frequency of the a allele, p, in the population is the square root of the disease frequency, that is, $p = (\pi)^{\frac{1}{2}}$. On the other hand, if a disease is dominant and caused by the presence of the A allele, with allele frequency p, then the disease frequency in the population is given by $\pi = p^2 + 2p(1 - p)$, or $p = 1 - \sqrt{1 - \pi}$.

The probability that a relative of a proband expresses a genetic trait can be worked out using the ITO-matrix method of Li and Sacks [13], and a selection of such values in terms of gene frequencies p and q (where $q = 1 - p$) is given in Table 2 for a single-locus trait and a variety of degrees of genetic relationship. For example, consider a recessive disease with frequency $\pi = 0.05$ in the population and hence allele frequency $p = 0.2236$. From Table 2, the risk to an offspring of an affected parent is simply $p = 0.2236$, whereas that to an offspring of an unaffected control is given by $p^2/(1 + p) = 0.0409$. Consequently, the value of RR is the ratio of these or 5.47, and $OR = 6.75$.

Familial risks are functions of the allele frequency in the population, because a relative can inherit the high-risk allele other than by direct descent from the proband. For

TABLE 3. RELATIVE RISKS* FOR SELECTED PAIRS OF RELATIVES, FOR DOMINANT AND RECESSIVE DISORDERS, AT VARIOUS LEVELS OF THE DISEASE RATE IN THE POPULATION (PURELY GENETIC DISORDER)

	Dominant gene				Recessive gene		
Disease rate (π)	Parent–offspring	Sib–sib	Uncle–nephew**		Parent–offspring	Sib–sib	Uncle–nephew**
0.001	1000.50	1000.75	334.22	*RR*	32.62	362.15	16.83
	2001.50	2002.25	445.70	*OR*	33.66	493.07	17.09
	500.38	500.44	250.69	*PRR*	31.62	266.06	16.31
0.010	100.50	100.75	34.22	*RR*	11.00	42.94	6.05
	201.50	202.26	45.71	*OR*	12.11	61.12	6.34
	50.38	50.44	25.69	*PRR*	10.00	30.25	5.50
0.100	10.49	10.74	4.22	*RR*	4.16	6.88	2.74
	21.54	22.36	5.72	*OR*	5.62	11.37	3.20
	5.38	5.44	3.19	*PRR*	3.16	4.33	2.08

*RR = Relative Risk, OR = Odds Ratio, PRR = Population Relative Risk; all defined in text. **Equivalent for any pair of aunt/uncle–nephew/niece.

example, in a dominantly inherited trait, the offspring of an affected heterozygote father may be affected be receiving the disease allele only from the mother, and her probability of bearing that allele is a function of its frequency. In general, high-risk alleles for chronic disease will be found to be rather rare, say $p < 0.10$ in the population, so that the absolute risk to relatives of unaffected probands would be small, but this may not always be the case and cannot be assumed.

In Table 3 we present numerical values for RR and OR for selected frequencies of a purely genetic disease, and for three types of close relatives. A third measure, PRR, is given and will be discussed later. The rarer the disease allele and the closer the degree of relationship, the larger is the relative risk. This is true, for example, in the case of bilateral retinoblastoma, in some families a heritable form of childhood retinal tumor [14]. Diseases like this are so rare that if they were not genetic the chance occurrence of even just two cases in the same sibship would be only one in several million families. Familial clustering of such conditions, therefore, can be striking, and relative risks for such rare diseases are extremely large—but because they are so rare they may not require relative risk computations to show their familiality.

On the other hand, as can be seen from Table 3, if an allele is more common relative risks can become small, because with a high gene frequency the probability that a relative inherits the disease allele from some relative other than the proband is not trivial in comparison with the probability of direct inheritance from the proband. An example of this is sickle cell anemia. The sickle-cell hemoglobin allele has a frequency which can be as high as about $p = 0.15$ in adult West Africans [12]. Among sibs of cases in such a population the expected value of RR would only be about 6.0, based on Table 2, yet this is purely genetic disease and the allele frequency is high.

Sampling considerations

We have considered the expected risk to a relative of an affected proband as compared to that of an unaffected proband. In case–control data, sampling considerations may place strong constraints on the ability to make reliable inference about genetic factors because of the nature of the statistics RR and OR. Sample values of these measures, since they are ratios of random variables and have skewed sampling distributions, are not necessarily good estimators of the underlying population values if sample sizes, especially the number of controls, are small or if the disease is rare in the population [2, 15]. In fact, the expected values of RR and OR are undefined, since for all sample sizes there is a positive probability of finding no affected relatives in the controls, that is (from Table 1), of n_{21} being equal to zero (or of all relatives of probands affected, in the case of OR, i.e. $n_{12} = 0$). The probability of n_{21} being zero is expecially high if P_2 and/or N_2 are small.

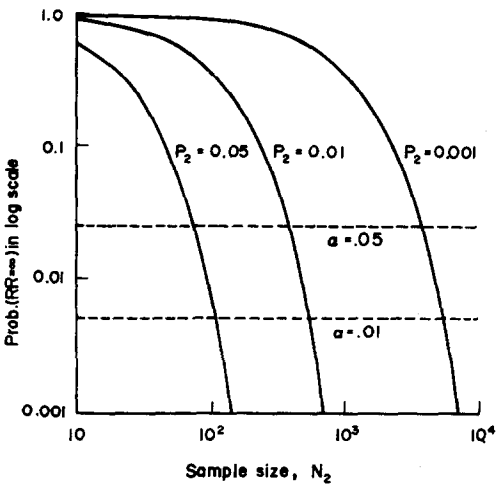

FIG. 1. Probability of observing no affected control relatives in a case–control study with N_2 control relatives sampled, for various values of the underlying probability of affected control relatives in the population, P_2. Dotted lines show sample sizes required to guarantee that upper limit of the $1 - \alpha$ confidence interval will be finite. Vertical axis, in logarithmic scale, is probability that the Relative Risk, RR, is infinite, due to $n_{21} = 0$ in the sample.

It has often occurred in practice that samples of data on relatives of controls contained very few (or no) cases of a disorder. It seems probable to us that many times such results have not been reported in terms of relative risks, possibly because the investigator considered the sampling design or its outcome to be faulty or uninformative, or because the association was taken to be sufficient proof of causation (see [3]). In the remaining cases with small P_2 and/or N_2, when a finite relative risk is observed and reported as such, the estimate of the relative risk in that case will often be smaller than the underlying population value. This negative bias in the estimation of the true RR or OR from small samples is strictly an artifact of inadequate samples, but the literature is replete with examples and this may tend to give an erroneous impression that the true relative risks are very small. Yet, when the frequencies of predisposing alleles are small and risk to control relatives consequently least, the true relative risks are largest, and these larger values may be missed or under-estimated because the absolute risks are so small as to require very large samples.

These sampling problems make it important not simply to present a sample value of the point estimated or RR or OR but to present its confidence interval as well. There are several large-sample methods for constructing confidence intervals from 2×2 tables by logarithmic transformation or to provide point estimators using Yates' continuity correction, etc. but the adequacy of the approximation with small P_2 and N_2 is still a matter of debate (e.g. [2, 16–17]). Complete enumeration of the sampling distribution is the safest way to study the small sample (small N_2, P_2) properties of RR and OR, because the approximate methods may be misleading. No matter how it is estimated, if the confidence interval is very wide, or if it does not have a finite upper bound (e.g. if the probability that $n_{12} = 0$ exceeds $\alpha/2$, for the $100(1 - \alpha)\%$ confidence interval), the observed relative risk has little practical value as a point estimate, even though it may indicate whether or not the measure is statistically significant.

To indicate the minimal sample sizes needed to ensure a finite upper bound to the 95 and 99% confidence intervals for RR, we present in Fig. 1 some values of the probability of observing $n_{21} = 0$ for particular sample sizes N_2 and a selection of P_2 values (note that $RR < 1/P_2$). OR can be undefined also if $n_{12} = 0$, if all relatives of cases are affected, but as this is unlikely to occur in a genetic context, we have ignored this circumstance. One should obtain samples of relatives considerably larger than these limits in order to get a narrow enough confidence interval to be useful in practice.

We recommend using the confidence interval in addition to the point estimate of relative risk because of the problems just discussed [18]. No careful investigator will use inadequate samples by design, but this can nonetheless occur if data must be stratified in several ways, such as by age, sex, types of relatives. The problem of adequate sample sizes after stratification may occur when family data are included in what is basically a search for exposure risks in a cohort, for example, of workers in a particular industry. When the resulting disease is rare, a chance occurrence of one or two high-risk families in those workers might greatly inflate the apparent exposure risk, so that family history must be included, but the expected risk to controls might be too low to ensure detection.

There is another restriction on the type and the size of the samples. Often, data on multiple relatives per case/control are gathered in studies of family risk history, for example multiple sibs, several offspring, etc. So far as testing the null hypothesis of no-association is concerned, there is no problem with this kind of data. But if the null hypothesis is rejected, then the accurate estimation of the relative risks and their standard errors becomes virtually impossible from data on multiple relatives, because of the dependency of these multiple observations under any genetic model. In general, unless complete pedigrees are gathered and the data subjected to formal genetic analytic methods [19–21], *only one relative per proband* can be used for estimation, even though this puts restrictions on sample sizes. In some situations, where family history data gathered without regard to probands' disease status may be available, there are some simplified methods which may be applied at least to estimate the power of tests to detect a given level of excess risk [22–24], and in this case the data may include multiple relatives.

2. MORE COMPLEX ETIOLOGY: INCOMPLETE PENETRANCE, SPORADIC CASES

For most chronic diseases it is likely that the majority of *cases* one would ascertain will be of environmental origin (not involving the inheritance of susceptibility alleles), even though such alleles do exist in the population. Such cases will not be familial unless there is shared family environmental exposure (which is very difficult to distinguish from genetic risk). We will refer to these cases as 'sporadic' cases, because that is how they appear in family data. In human genetics terminology they are called 'phenocopies'. On the other hand, some bearers of susceptibility alleles may not manifest the disease at the time they are observed; this failure to express a phenotype is common with medical disorders and is known as 'incomplete penetrance'. One result of incomplete penetrance is that some *controls* one might ascertain could be the bearers of high-risk alleles even though unaffected at time of ascertainment. Incomplete penetrance can be due to the mollifying effect of other gene loci, manifestation of the trait only at a subclinical level, lack of exposure to some environmental stimulus, delayed age of onset, and so on.

The result of sporadic cases and incomplete penetrance in controls is that dividing family data according to cases and controls does not provide the kind of sharp division in exposure (of relatives) which is the rationale for the case-control or cohort-exposure approaches. The resulting relative risks should be carefully interpreted in terms of the *strength* of the genetic effect. To avoid some of the estimation problems occasioned by the use of controls, and to gain a more intuitively interpretable result, one can use what we will here term the Population Relative Risk, *PRR*, defined as

$$PRR = \frac{\text{Prob (Relative is affected | Proband is affected)}}{\text{Prob (Random member of population is affected)}} = \frac{P_1}{\pi}. \tag{1}$$

The meaning of *PRR* is clear and relevant to chronic disease epidemiology: it is the excess risk of the relative of an affected individual as compared to that of a random member of the population. This is a straightforward way to express familial risk. A sample of controls is not used, but rather the value of the population risk, π, obtained from some secondary source such as national health statistics (censuses, diseases regis-

tries, etc.) for a group demographically like the cases and their relatives, is used. Data of these kinds are often available, and when based on complete enumeration, π can be assumed known without any sampling error. Even when the secondary source is not as exhaustive as a census or registry, it may be based on such large samples that there is effectively no sampling error. When this can be assumed, the sample $\widehat{PRR}$

$$\widehat{PRR} = \frac{\hat{P}_1}{\pi} = \frac{n_{11}/N_1}{\pi} \tag{2}$$

is an unbiased estimator of the underlying *PRR* in the population, and provides the kind of 'observed/expected' test desired of family history data. The confidence interval can be computed directly from the variance, which is

$$\text{Var}(\widehat{PRR}) = \frac{\hat{P}_1(1 - \hat{P}_1)}{N_1\pi^2}, \tag{3}$$

because if samples are large enough the distribution of *PRR* is asymptotically normal. The test of significance of association of risk is thus straightforward:

$$\chi_1^2 = \frac{N_1(\hat{P}_1 - \pi)^2}{\pi(1 - \pi)}. \tag{4}$$

In Table 3 we have provided values of *PRR* corresponding to the *RR* and *OR* which would arise under the same purely genetic conditions.

The *PRR* is a form of indirectly standardized risk ratio, such as the standardized mortality or standardized incidence ratio (SMR or SIR), comparing a chosen subset of a population to the population of which it is a constituent. The population risk includes both the high and low risk subgroups, so that *PRR* is smaller than measures of risk which are based on clear exposure or risk divisions, but when one cannot construct samples with such a well-defined exposure it is reasonable to use such a measure, which at least is based on a meaningful comparison group (the whole population). *PRR* has been used by geneticists in the past both to detect familial risk and to illustrate theoretically the expected effects of genetic risk. Penrose derived tables similar to our Table 3, allowing for incomplete penetrance [25]. Edwards [26] and others [27, 28] have used *PRR* or similar measures to show the differences in expectation arising from different modes of inheritance such a dominant, recessive, and multilocus. It is difficult to distinguish between single locus traits and multilocus traits which are expressed dichotomously (affected/unaffected) only when a genetic threshold has been reached [26], and many diseases produce *PRR* values which are compatible with a mixture of environmental and genetic effects [29]. Krüger [12, 30] has given a series of nomograms which can be used, with special kinds of family data, to discriminate among different modes of inheritance. Although many diseases may be multilocus in etiology, there is good evidence, for a wide range of chronic diseases, that only one major locus is important, so that in this paper we are only interested in the single-locus case. Conclusions in regard to multilocus conditions would be similar to those given here.

Even without the complicating factors of incomplete penetrance and the occurrence of sporadic cases, Table 3 shows that *PRR* has smaller values than do the other measures. When there is incomplete penetrance and sporadics, all the measures are reduced. Incomplete penetrance usually reduces the numerator of *PRR*, and sporadic cases disproportionately increase the denominator, both effects lowering the value of the index. Table 4 provides a selection of *PRR* values for sib–sib pairs and a range of disease frequency, penetrance, and sporadic rates to illustrate the effects which can occur. The relationship among the various parameters is complex; for simplicity we avoid these formulae here as they are available elsewhere [31]. For comparative purposes, values are in Table 4

TABLE 4. VALUES OF POPULATION RELATIVE RISK (PRR) IN SIB–SIB PAIRS WITH INCOMPLETE PENETRANCE ANS SPORADIC CASES, FOR SEVERAL VALUES OF DISEASE FREQUENCY

Disease frequency (π)	Proportion of sporadics	Dominant Penetrance fraction 1	0.75	0.5	Recessive
0.001	0	500.44	373.745	250.750	266.061
	0.5	248.671	189.218	121.283	131.030
	0.99	5.990	3.878	2.338	3.503
0.01	0	50.443	37.925	25.750	30.250
	0.5	25.524	19.925	13.564	14.725
	0.99	1.490	1.294	1.147	1.250
0.1	0	5.441	4.283	3.250	4.331
	0.5	3.210	2.774	2.292	2.324
	0.99	1.040	1.027	1.016	1.043

Calculations based on [31], see text. For this table, disease frequency, π, taken as independent variable, gene frequency then computed for each mode of inheritance for specified sporadic fraction, penetrance 1.0. Incomplete penetrance not considered for recessive case. Within section specified by a value of π, actual disease frequency varies slightly, as a consequence of interdependency of variables, with changes in sporadic and penetrance fractions.

corresponding to the purely genetic conditions of Table 3 (penetrance = 1.0, sporadic rate = 0.0). We have not discussed details of incomplete penetrance in this paper, and as the concept is somewhat involved in relation to recessive traits we only present fully penetrant recessive *PRR* values. Clearly, the more common the disease and the higher the fraction of sporadics (i.e. environmental and stochastic cases), the smaller is *PRR*, and *PRR* values of 2–5 are not difficult to obtain. Note also that under these conditions one cannot expect to find cases arising in sibships in the classic Mendelian proportions (such as $\frac{1}{4}$ or $\frac{1}{2}$), and such proportions should *not* be considered as the required signposts of genetic risk.

An illustration of the more complex situation can be found in adult-onset diabetes mellitus. The lifetime probability of diabetes in Western Europeans living to old age is about 5%. Although there are clearly some nongenetic causal factors related to obesity, diet, and other aspects of lifestyle, there is persistent evidence for additional familial aggregation, and the child of one affected parent has about a 20% lifetime risk [11]. This yields a population value of *PRR* of about 4.0 in such parent/child pairs. The risk of diabetes in Mexican–Americans is much higher than 5%, perhaps as high as 25–50% in those living to old age, based on preliminary work done at this center. Risk in Amerindians in the southwestern U.S. is similarly elevated. The shared ancestry of these groups, along with evidence from family data, provide very strong indication that a sizeable fraction of cases in these populations may be of genetic origin. If every case were genetic, one might expect to find 50% or more children of affected probands to be affected (Mendelian proportions) but if the value of π is about 0.25 or more, *PRR* will be substantially less than 4.0—about 2.0 or even less. This is an artifact of *PRR*, and in no way indicates a minimal involvement of genetic factors in the disease. The effect of high prevalance or incidence in the population in reducing relative risks has been noted before [32].

Effects of differential age of onset

Almost by definition, a degenerative or chronic disease is one which has variable age of onset, and risk which increases with age. It can be important to decide whether, in a given case, it is better to express risk in terms of incidence, prevalence, or cumulative incidence. Often, studies are mixed in this regard. Cases may be ascertained at incidence, that is, individuals affected *at* their age of observation, whereas relatives may only be classified as to prevalence, whether or not affected *by* their last observed age. Studies will rarely be comparable in terms of the numbers, sexes and ages of the relatives obtained so

that comparing relative risks from more than one investigation into the same disease may be difficult. Incidence per person–year at risk might be the best measure to use in dealing with age–onset problems, cumulating such incidence for a given age range. This may depend to some extent on the nature of the disease, for example, as to whether or not the disease can recur more than once in the same individual and how those who have recovered are to be classified.

The absolute risk of onset changes with age by several orders of magnitude for most degenerative diseases, so that the sampling variance of a measure like *PRR* is sensitive to age (equation 3). Thus, if a pooled significance test is desired, the pooled variance of the test statistic must be a weighted average of the variances of the age subclasses. A Mantel–Haenszel test [33, 34] against population risk, can take into account differences in risk between age strata as well as stratification by degree of relationship. Such a test assumes that the direction of risk difference is the same for all strata, but in relation to age this may not always be the case with genetic risk and the test may consequently fail to discriminate between a true circumstance of no-association and an association manifest only in certain subsets of the data. In addition, the Mantel–Haenszel test is used when it can be assumed that the magnitude, as well as the direction, or risk difference is constant across all strata, which suggests that a separate test for each type of relationship would be better than a single pooled test.

For many degenerative diseases the effect of familial factors seems to be to raise the risk at young adult ages, that is, to shift the age–onset curve toward younger years. It is not clear whether genetic risk remains in excess at older ages. Often, the clinical pathology of the disease is basically the same, so far as one can tell, in both familial and sporadic cases and the *only* discernable effect of high risk alleles is to lower the ages of onset. The best approach to detect genetic risk when this may be occurring is to stratify data according to the ages of *both* probands *and* relatives. The age and disease status of a proband affect the probability that he/she is the bearer of a high-risk allele, which in turn affects the probability that the relative received such an allele from the proband. Similarly, the relative's age and disease status are informative as to whether an allele was in fact inherited. A young affected individual is more likely to be the bearer of a genetic risk factor than one affected at old age, and the child of the former will have a higher expected risk than a child of the latter. If family history data are stratified in this manner, genetically-based onset effects produce *PRR* greater than 1.0 only in some subsets of the data, revealing their existence. Two-way stratification may place constraints on sample sizes, but this should be done at least by dividing cases and their relatives each into two broad age classes, such as below and above age 50. Failing to stratify in this way may simply obscure the evidence for familial risk. Technically speaking, in fact, one might desire data on the age and affection status of unascertained relatives as being relevant to the problem. For example, if differential mortality or infertility due to a genetic disease selects against susceptibility alleles in parents, the risk in sib sets may be better assessed if they are classified according to the ages of their parents at the birth of those children. Breast cancer, vulnerability to ovarian cysts, and similar conditions are examples where this might be important; risk would be likely to show up in sib–sib data mostly in those born to young mothers because women surviving and able to bear children at older ages are less likely to carry high risk alleles.

An example of the age effect can be found in the condition known as Familial Polyposis Coli (FPC). This is a single-locus dominant genetic disorder [7, 10]. Individuals bearing the FPC allele develop numerous polyps in the colon during young adult years and in about 80% or more of such individuals one or more of the polyps is transformed to become carcinoma of the colon by age 50. In the U.S. at large, the risk of colon cancer by age 50 is only about 0.002, so that a person with early onset of colon cancer has a substantial probability of being a victim of FPC (or one of several similar syndromes) and the *PRR* for a child of such a person is very much above the usual 2–5 fold relative risk observed in general colon cancer data [10]. In data on incidence per person–year of risk classified according to age, the expected *PRR* declines as the ages of the children of

an early-onset parent increase, because the older these unaffected children are the lower is the chance that they inherited the FPC allele. The older a proband is at onset, the lower the chance that he/she has FPC and the lower the value of *PRR* for the children, to the point that the children of a late-onset case can have *PRR* slightly less than 1.0, that is, less than the risk to a randomly chosen member of the population of the same age [35–37]. An old, unaffected offspring even of a known FPC victim can have *PRR* less than 1.0 when ages are properly accounted for. The precursor lesion for FPC (polyps) is well known, so that individuals can be assigned to their proper risk group via colonoscopic examination. But there may be many similar kinds of genetic subsets of otherwise common diseases which no not have a currently known precursor state, and it would be easy to miss the familial excess risk in family history data not stratified by the ages of both probands and relatives.

The use of age-stratified relative risks has been successful in identifying high-risk individuals in whom specific genetic factors are probably operating. This has been the case in regard to breast cancer. Many family history studies showed relative risks on the order of about 2–3 [5, 38–41]. By age subdivision, Anderson [42] identified families whose relative risks were very greatly elevated (estimated to be about 50); these were bilateral, premenopausal, multiply-affected nuclear families. Large-scale studies of Utah Mormon pedigrees have corroborated both the familiality of breast cancer and the nature of the high-risk subsets [43, 44], and recently several workers using similar criteria to identify probands have claimed that a single gene, located on chromosome 10, may be responsible for at least some cases of this kind of early-onset breast cancer [45, 46].

DISCUSSION

A relative risk is a labile index whose value depends on the type and ages of the individuals from whom is was derived. Apart from its use to detect risk excess, there may not be a correspondence of relative risk values from different studies of the same disease, and clearly there is no single or "true" familial relative risk in a population. Much the same could be said in regard to environmental relative risk, except that in the genetic case there is less design control over the degree of exposure in those sampled. Epidemiologists also seem to be less sensitive to these issues in relation to familial factors.

What is one really trying to estimate from a sample relative risk estimate? Probably, this is some form of lifetime risk of expression of a disease in the relative of an affected proband as compared to that in the population or to relatives of unaffected controls. This would make intuitive sense both epidemiologically and to the geneticist advising family members of their risks, but such data are usually not available from casual family studies. Furthermore, one must decide whether "lifetime" should refer to the natural lifetimes of actual individuals, in which case one must consider the dynamics of competing causes of mortality, or to the risk to an individual *if* he/she lives to a specified old age, that is, the risk should no other cause intervene. Even here, however, one would have to decide whether such lifetime relative risks should be conditioned on the age of onset in the proband, and again it seems that no single index will be fully satisfactory. These things are rarely considered.

There are two additional sources of error to which family data are particularly prone and which can affect relative risk analysis. One is ascertainment bias. There are many ways in which affected family members can come to the attention of the investigator disproportionately as compared to unaffected members (e.g. [12]). Multiply-affected families may be more likely to show up in referral clinics, and there is great recall bias in the use of anamnestic data alone, on which many studies rely. Affected probands may have heightened awareness of other cases of the same disease in their family, relative to unaffected probands, which may lead to an upward bias; on the other hand for social or other reasons relatives may not reveal the true nature of a diagnosis, leading to a downward bias. One should verify, for example, whether a relative said to have had breast cancer actually had benign breast disease, or vice versa.

The second source of error is pedigree error. Careful pedigree studies using biochemical marker genes usually detect a sizeable percentage of unreported nonpaternity. This may be due to unknown nonpaternity, unreported illegitmacy, adoptions, and so on. Pedigree errors usually result in an underestimate of familial risk when genes are involved because they substitute random individuals for related ones [36]. Family history based on interview or questionnaire is vulnerable to this source of error, although it takes a considerable effort, including genetic studies of blood samples, to obtain a good estimate of the extent to which it has occurred in a particular study.

We have reviewed many difficulties with the use of casual family history data and showed that small relative risks can result in a number of ways and that relative risks can in any case be difficult to interpret. Does this mean that use should not be made of such data? To the contrary, we would argue that it is very important to include family history, with all of its pitfalls, in any reasonably comprehensive risk-factor survey. Furthermore, it is a realistic fact that a casual family history is the only economically practicable way to examine family risk for large numbers of diseases, at least initially. We would only stress three major cautions in interpreting results from such work: (1) small relative risks, of the order of 2, should not be overlooked [35]; (2) small relative risks indicate, generally, a large fraction of cases are sporadic but do *not* indicate that there is little genetic involvement; and (3) casual family history data should only be used to test the null hypothesis and should not be expected to be very informative about the nature of genetic factors, which require follow-up studies. What are the benefits to following up relative risks which are greater than 1.0, even if small? We believe that they are basically three.

First, the *epidemiology* of genetics of a disease is a function of parameters like gene frequency which underlie the process, and when preliminary studies have indicated that there is excess familial risk these parameters can be estimated and the mode of inheritance of risk better understood. The method of choice for doing so is known as complex segregation analysis. This is an extension of basic principles of Mendelian segregation to cover a range of complexities including sporadics and incomplete penetrance (e.g. [19, 20]). The method requires family data which have been collected according to certain principles for avoiding ascertainment bias [47–49]. Segregation analysis does not involve concepts like strength-of-effect, but rather is a means of testing the goodness of fit of a specific genetic model to data and of estimating the underlying parameters. Results from different studies are directly comparable, and age effects can be taken into account through age-onset functions assigned to different genotypes by the model. High risk individuals can be identified within the family structures.

The second reason has to do with *etiology*. If susceptibility is due to a very common allele, most people may be susceptible, as might be the case with regard to some infectious disease; one might wish to look at low risk individuals in order to understand the immunogenetics of their resistance. More often, however, only a minority of individuals will be predisposed and it can be important to identify them to understand the genetic mechanisms of their risk. This may not simply lead to an understanding of the genetics of the disease in those individuals but also in the population at large. Many chronic diseases seem ultimately to be the result of some genetic event(s): the incorporation of a viral gene into the genome, the stochastic occurrence of a mutation producing a clone of autoantibodies, the accumulation of somatic mutations as the result of exposure to environmental mutagens, etc. Solid adult carcinomas are generally thought to be the result of the latter kind of mutational process. Although sporadic cases may be the result of such changes in somatic tissue, occasionally the *same* mutational events may occur in germ line cells and be inheritable in the bearers of such mutations and their descendants in the population [50]. High risk families provide an opportunity to see the alternative alleles segregating, with a consequent opportunity to perform genetic linkage studies to locate such alleles on their chromosome or, with the genetic technology rapidly becoming available, to identify the locus or its gene products themselves. Such identification is an important strategy in contemporary genetic epidemiology.

The final reason for following up family risk studies is related to treatment. The identification of a high risk subset of individuals can be important to an understanding of the course of the disease and to assigning risk to individuals based on prodromal symptoms which may be identified through studying them. These individuals would benefit from genetic counseling to advise them of their risk and that to their family members and to suggest prophylactic measures they may take; they may also benefit from screening for early detection. Although the genetically predisposed may comprise only a small proportion of all cases of a disease in the population they may be a numerically important group.

Acknowledgements—This work was accomplished with financial support from the National Cancer Institute, Grant CA 19311, the National Institute of Aging, Grant AG 01028, the National Institute of General Medical Sciences, Grant KO4 00230, and Department of Energy, DE-AC 02-76EV 02828, which support is gratefully acknowledged. We also would like to thank Dr William J. Schull and Dr J. Fred Annegers for their help in improving early versions of the manuscript. This is Demographic Epidemiology of Aging and Disease, Paper No. 13.

REFERENCES

1. Fleiss J: **Statistical Methods for Rates and Proportions.** New York: Wiley, 1973
2. Gart J: Statistical analysis of the relative risk. **Envir Hlth Persp** 32: 157–167, 1979
3. Fienberg S: **The Analysis of Cross-Classified Categorical Data.** Cambridge, MA: MIT Press, 1977, p. 18
4. McKusic V: **Mendelian Inheritance in Man.** 5th Edn Baltimore: Johns Hopkins University Press, 1978
5. Lilienfeld AM: Epidemiological concepts applied to studies of chronic diseases. In **Genetics and the Epidemiology of Chronic Diseases.** Neel J, Shaw M, and Schull W (Eds). Washington, D.C.: National Institute of Health, 1965, pp. 87–102
6. Tokuhata GK, Lilienfeld AM: Familial aggregation of lung cancer in humans. **J Nat Cancer Inst** 30: 289–312, 1963
7. Lipkin M, Sherlock P, Decosse J: Risk factors and preventive measures in the control of cancer of the large intestine. **Curr Probl. Cancer** 4. Chicago: Yearbook Medical, 1980
8. Lynch H, Lynch P, Albano W, Edney J, Organ C, Lynch J: Hereditary cancer: Ascertainment and management. **Ca—A Cancer Journal for Clinicans** 29: 216–232, 1979
9. Lovett E: Family studies in cancer of the colon and rectum. **Brit J Surg** 63: 13–18, 1976
10. Correa P, Haenszel W: The epidemiology of large-bowel cancer. **Adv Cancer Res** 26: 1–141, 1978
11. Stevenson A, Davison B, Oakes M: **Genetic Counseling.** Philadelphia: Lippincott, 1976, p. 259
12. Vogel F, Motulsky A: **Human Genetics.** Heidelberg: Springer-Verlag, 1980
13. Li C, Sacks L: The derivation of joint distribution and correlation between relatives by the use of stochastic matrices. **Biometrics** 10: 347–360, 1954
14. Vogel F: Genetics of retinoblastoma. **Hum Genet** 52: 1–54, 1979
15. Yanagawa T: Designing case-control studies. **Envir Hlth Persp** 32: 143–156, 1979
16. Fleiss J: Confidence intervals for the odds ratio in case–control studies: The state of the art. **J Chron Dis** 32: 69–77, 1979
17. Haber M: A comparison of some continuity corrections for the chi-squared test on 2 × 2 tables. **J Am Stat Assoc** 75: 510–515, 1980
18. Cole P: The evolving case-control study. **J Chron Dis** 32: 15–27, 1979
19. Elston R: Segregation analysis. In **Current Developments in Anthropological Genetics.** Mielke J, Crawford M (Eds). New York: Plenum, 1980, pp. 327–354
20. Morton N, Rao D: Casual analysis of family resemblance. In **Genetic Analysis of Common Diseases: Application to Predictive Factors in Coronary Disease.** Sing C, Skolnick M (Eds). New York: Alan R. Liss, 1979, pp. 431–452.
21. Schull W, Weiss K: Genetic epidemiology: four strategies. **Epid Rev** 2: 1–18, 1980
22. Chakraborty R, Weiss K, Schull W: A test for the randomness of the occurrence of a disease trait in familial or other similar ordered sequences of epidemiological data. **Proc Natl Acad Sci** 77: 2974–2978, 1980
23. Smouse P, Weiss K, Chakraborty R: A simple test for the aggregation of disease occurrence in genealogical data. **Hum Hered** 31: 334–338, 1981
24. Carmelli D, Karlin S, Williams RR: A class of indices to assess major-gene versus polygenic inheritance of distributive variables. In **Genetic Analysis of Common Diseases: Applications to Predictive Factors in Coronary Disease.** Sing C, Skolnick M (Eds). New York: Alan R Liss, 1979, pp. 259–270
25. Penrose LS: The genetical background of common diseases. **Acta Genet** 4: 257–265, 1953
26. Edwards JH: The simulation of Mendelism. **Acta Genet Stat Med** 10: 63–70, 1960
27. Falconer DS: The inheritance of liability to certain diseases estimated from the incidence among relatives. **Ann Hum Genet** 29: 51–76, 1965
28. Cavalli-Sforza LL, Bodmer WF: **The Genetics of Human Populations.** San Francisco: W. Freeman, 1971
29. Newcombe HB: Discussion. In **Second International Conference on Congenital Malformations.** New York: International Medical Congress, 1964, pp. 345–347
30. Kruger J: Zur unterscheidung zwischen multifaktoriellem erbgang mit schwellen werteffekt und einfachem dialelem erbgang. **Hum Genet** 17: 181–252, 1973
31. Majumder PP, Chakraborty R, Weiss K: Relative risks of common diseases in the presence of incomplete penetrance and sporadics. (Submitted)
32. Cornfield J: A statistical problem arising from retrospective studies. **Proc. 3rd Berkeley Symp** IV: 135–148, 1956

33. Mantel N, Haenszel W: Statistical aspects of the analysis of data from retrospective studies of disease. **J Natl Cancer Inst** 22: 719–748, 1959
34. Elandt-Johnson RC, Johnson NL: **Survival Models and Data Analysis.** New York: Wiley, 1980
35. Peto J: Genetic predisposition to cancer. In **Banbury Report No. 4: Cancer Incidence in Defined Populations.** Cairns J, Lyon J, Skolnick M (Eds). Cold Spring Harbor, New York: Cold Spring Harbor Laboratory, 1980, pp. 203–213
36. Chakraborty R, Weiss K, Ward R: Evaluation of risks from the correlation between relatives. **J Med Anthrop** Summer: 397–413, 1980
37. Weiss K, Chakraborty R, Schull W, Rossmann D, Norton S: The Laredo Epidemiology Project. In **Banbury Report No. 4: Cancer Incidence in Defined Populations.** Cairns J, Lyon L, Skolnick M (Eds). Cold Spring Harbor, New York: Cold Spring Harbor Laboratory, 1980, pp. 267–284
38. Macklin M: Genetic considerations in human breast and gastric cancer. In **Genetics and Cancer.** Austin, Texas: University of Texas Press, 1959, pp. 408–425.
39. Oliver C: Genetic studies of families with high cancer incidence. In **Genetics and Cancer.** Austin, Texas: University of Texas Press, 1959, pp. 426–438
40. Albert S, Child M: Familial cancer in the general population. **Cancer** 40: 1674–1679
41. Henderson B, Powell D, Rosario I, Keys C, Hanisch R, Young M, Casagrande J, Gerkins V, Pike M: An epidemiologic study of breast cancer. **J Natl Cancer Inst** 53: 609–614, 1974
42. Anderson D: Genetic study of breast cancer: Identification of a high risk group. **Cancer** 34: 1090–1096, 1974
43. Hill J: A survey of cancer sites by kinship in the Utah Mormon population. In **Banbury Report No. 4: Cancer Incidence in Defined Populations.** Cairns J, Lyon L, Skolnick (Eds). Cold Spring Harbor, New York: Cold Spring Harbor Laboratory, 1980, pp. 299–316
44. Hill J, Carmelli D, Gardner E, Skolnick M: Likelihood analysis of breast cancer predisposition in a Mormon pedigree. In **Genetic Epidemiology.** Morton N, Chung C (Eds). New York: Academic Press, 1979, p. 304
45. King M, Go C, Elston R, Lynch H, Petrakis N: Allele-increasing susceptibility to human breast cancer may be linked to the glutamate-pyruvase transaminase locus. **Science** 208: 406–408, 1980
46. King M, Bishop T: Preliminary analysis for linkage of glutamate–pyruvase transaminase and breast cancer susceptibility in a Mormon kindred. In **Banbury Report No. 4: Cancer Incidence in Defined Populations.** Cairns J, Lyon L, Skolnick M (Eds). Cold Spring Harbor, New York: Cold Spring Harbor Laboratory, 1980, pp. 379–385
47. Elston R, Sobel E: Sampling considerations in the gathering and analysis of pedigree data. **Am J Hum Genet** 31: 62–69, 1979
48. Cannings C, Thompson EA: Ascertainment in the sequential sampling of pedigrees. **Clin Genet** 12: 208–212, 1977
49. Thompson EA, Cannings C: Sampling schemes and ascertainment. In **Genetic Analysis of Common Diseases: Applications to Predictive Factors in Coronary Disease.** Sing C, Skolnick M (Eds). New York: Alan R. Liss, 1979, pp. 363–382
50. Knudson AG: Genetics and the etiology of human cancer. In **Advances in Human Cancer.** Harris H, Hirschhorn K (Eds). New York: Plenum, 1977 (Vol. 8), pp. 1–66
51. Breslow NE, Day NE: **Statistical Methods in Cancer Research.** Vol. 1. The Analysis of Case-Control Studies. Lyon, France: International Agency for Research on Cancer, 1980