Statistical Methods in Medical Research 2000; **9**: 543–562

Statistical designs for familial aggregation

Kung-Yee Liang Department of Biostatistics and **Terri H Beaty** Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, Maryland, USA

In the past two decades, it has become increasingly clear that genetic factors contribute to the aetiology of many common diseases including cancers, coronary disease, allergy and psychiatric disorders. While one goal of genetic epidemiological studies is to locate susceptibility genes for these complex diseases, it is important that strong evidence of familial aggregation be established at an early stage of research. In this paper, we discuss several study designs useful to address some issues such as (1) detecting familial aggregation, (2) testing for gene-environment interaction, (3) identifying homogeneous subgroups and (4) measuring magnitude and patterns of familial correlations. These designs include the conventional case-control design and the family case-control design. For each of these two study designs, we discuss analytical strategies as well as their strengths and weaknesses. Throughout, several examples from real studies are used for illustrative purposes.

1 Introduction

Many common diseases, including cancers, coronary heart disease, allergies and psychiatric disorders are known to be complex in aetiology since genetics and environmental factors contribute to the disease process. Genetic epidemiology is a relatively new field which utilizes the conventional epidemiologic designs and methods to explore the role genetic factors play in determining disease. While one goal is to locate susceptibility genes controlling risk to complex disease, a preliminary question in genetic epidemiology should be 'does the disease cluster in families?' More formally, the notion of familial aggregation or clustering amounts to a prevalence of disease among family members exceeding the expected for the general population. It is entirely possible that a disease having no genetic aetiology could also show evidence of familial aggregation or clustering, due to a shared exposure to an infectious disease or culturally transmitted risk factor. Nevertheless, strong empirical evidence of familial aggregation should be a prerequisite for further investigation into genetic mechanisms. In this paper, we discuss pros and cons of several analytical strategies commonly adopted to address the following related issues:

1) Does disease cluster in families and can such clustering be explained by genes or shared environment?
2) How does a specific gene, in conjunction with environmental factors, control risk of disease?
3) How can one identify genetically homogeneous subgroups of patients to enhance the chance of locating susceptibility genes?
4) How best to measure magnitudes and patterns of familial correlations?

Address for correspondence: K-Y Liang, Department of Biostatistics, Johns Hopkins University, 615 North Wolfe Street, Baltimore, MD 21205, USA. E-mail: kyliang@jhsph.edu

0962-2802(00)SM230RA

In the interest of brevity, we focus on two designs: the conventional case-control design and the family case-control design, although a number of other study designs are available.[1] Throughout, examples from several real studies will be used for illustrative purposes.

2 The conventional case-control design

2.1 Detecting familial aggregation

Considering question (1) above, a simple approach to detect familial aggregation is to compare the prevalence of a family history of the disease between cases and controls. This simple extension of the classic case-control design, commonly adopted by epidemiologists, helps to identify risk factors for either prevention or intervention purposes. Here the notion of positive family history is usually defined as the presence of disease in one or more first-degree relatives for either the cases or the controls. This information can be acquired by interviewing cases and controls themselves or through family informants. Thus if both cases and controls are drawn at random from their respective target populations, family history data can be simply expressed as in the 2×2 table as in Figure 1. This simple approach may be modified in the design stage so that the distribution of some important confounding variables are comparable between the sample of cases and controls. This matching includes either frequency matching or individual matching. In either case, the data can be summarized in a series of K tables, where K is the number of categories of a confounding variable in the frequency matching situation, or K is the number of cases when individual matching is adopted.

With K 2×2 tables one can use, for example, the Mantel–Haenszel[2] procedure for making inferences about the association between family history and case status. Here the parameter of interest would be the odds ratio (OR) defined for $i = 1, \ldots, K$, as

$$\mathrm{OR}_i = \frac{\Pr(\mathrm{PFH}|D, i\text{th stratum})/\Pr(\mathrm{NFH}|D, i\text{th stratum})}{\Pr(\mathrm{PFH}|\overline{D}, i\text{th stratum})/\Pr(\mathrm{NFH}|\overline{D}, i\text{th stratum})}$$

where D (or $\overline{D}$) denotes cases (or controls) and PFH (or NFH) for positive (or negative) family history. An odds ratio greater than one suggests a positive association between an individual's family history of the disease and his/her risk for the same disease. Depending on the magnitude of these OR, positive association may provide sufficiently strong evidence of familial aggregation to warrant further investigation

	Family history	
	Positive	Negative
Cases		
Controls		

Figure 1

into genetic mechanisms. As an illustration, consider the genetic epidemiologic study of chronic obstructive pulmonary disease (COPD) conducted by Cohen[3] in which 105 cases and 79 controls were sampled from the Johns Hopkins Hospital. The 2×2 table in Figure 2 indicates that 47.6% (= 50/105) of cases had a positive family history of impaired pulmonary function compared to 29.1% (= 23/79) of controls. This gives rise to an estimated log OR $= 0.79$ (with s.e. $= 0.315$) which is statistically significantly different than zero, i.e. the null hypothesis of OR= 1 is rejected.

It is interesting to point out that the variable 'family history' (FH) is different than the conventional risk factors observed on cases and controls in several regards. First, FH is not an attribute of cases and controls themselves, rather it depends on external factors including family size, biologic relationships of relatives to the index case, the age distribution of these relatives as well as the disease prevalence.[4] As a result, this variable is subject to misclassification. For example, even when the disease has no genetic aetiology, the probability of a case having a positive family history is

$$1 - (1 - \pi)^n$$

where π is the disease prevalence and n is the number of first-degree relatives. With $\pi = 0.05$, this proportion would be 0.19 when $n = 4$ and increases to 0.34 when $n = 8$. Meanwhile, a proportion of 0.34 is expected if the disease prevalence is 0.1 instead of 0.05 even if $n = 4$. Thus if the distribution of family sizes differs substantially between cases and controls, false estimates of this odds ratio may result. Another concern about the use of family history is the potential bias of information or recall. Depending on biologic relationship, number of relatives and other factors, the true disease status of relatives may be misreported, leading to further misclassification of the FH variable. Unlike the previous concerns, however, the degree of recall bias may well be differential between cases and controls. Consequently, the estimated odds ratio could either be attenuated or inflated and the magnitude of this discrepancy can be substantial.[4] This concern about potential recall bias may be alleviated, to some extent, by carefully choosing informants as the FH information is being gathered. For example, rather than only interviewing cases (or controls) directly, one may instead consider use of parents or spouses as informants (multiple informants may also be desirable). Indeed, in the situation where the cases (or controls) are deceased, this alternative becomes a necessity.

	Family history		
	Positive	Negative	
Cases	50	55	105
Controls	23	56	79
	73	111	184

Figure 2

An alternative to address concerns about varied family sizes, age distribution and biologic relationships across cases and controls is to create for each family a family history score.[4] This is accomplished by first deriving for each case or control an expected number of affected relatives

$$E = \sum_{j=1}^{n} E_j$$

where E_j is the expected risk to the jth relative based on risks from the general population considering age, gender and other demographic variables such as birth year. A family history score (FHS) is then defined as

$$(O - E)/E^{1/2} \tag{1}$$

the 'standardized' version of O, the observed number of affected among n relatives. Returning to the COPD study, the average scores for cases and controls are 0.383 (standard error (s.e.) $= 0.117$) and -0.006 (s.e. $=$ 0.107), respectively. A simple T-test reveals a significant difference in FHS between cases and controls, suggesting case relatives are at excess risk compared to control relatives, as measured by FHS.

2.2 Testing gene–environment interactions

This same case-control design can be further utilized to test for interactions between genes and environments. In the absence of knowledge regarding specific susceptibility genes, one may use either FH or FHS variable as a surrogate measure of 'genetic loading' or one can use markers in candidate genes. To test the hypothesis of interaction between environmental factors and genetic factors, such as family history (FH or FHS), one can consider the following logistic regression model, commonly adopted in case-control studies[5]

$$\text{logit Pr}(D|\text{FH(S)}, \text{ENV}) = \alpha + \beta_1 \text{FH(S)} + \beta_2\ \text{ENV} + \beta_3\ \text{FH(S)} \star \text{ENV} \tag{2}$$

where ENV stands for an observed environmental variable such as maternal smoking, etc. One can test the hypothesis of no interaction by examining the magnitude and direction of β_3, the coefficient for the interaction term. It is worth noting that this interaction, if it exists, acts in a multiplicative fashion under this model. Specifically, when comparing two individuals who differ by one unit in ENV, $E + 1$ versus E say, the odds ratio relating the disease to ENV for those with positive family history (PFH) is e^{β_3} times that for those without family history (NFH), i.e.

$$\text{OR}_{\text{PFH}}(E + 1, E) = \mathrm{e}^{\beta_2+\beta_3} \text{ and } \text{OR}_{\text{NFH}}(E + 1, E) = \mathrm{e}^{\beta_2}$$

In the situation where a candidate gene is observed, one can test for interaction between the marker genotype at the candidate gene and environment by applying (2) with FH(S) replaced by a dichotomous variable GEN which is 1 if the case (or control) carries the targeted allele(s) and 0 otherwise.

As an illustration, consider a case-control study on oral clefts in which 333 children born with oral clefts and 166 healthy infants were sampled.[6] A main objective of the study is to examine the association between oral cleft and a candidate gene, transforming factor alpha locus (TGFA), and the possible gene–environment interaction between TGFA and maternal smoking (MS). These data are summarized in Figure 3 in two 2 × 2 tables stratified by the maternal smoking status; here G (or $\overline{G}$) denotes the presence of C2 allele at the taq I polymorphism in TGFA. There is little evidence of association between oral cleft and TGFA ($\widehat{\mathrm{OR}} = 1.05$ and 1.07, respectively) whether the mother smokes or not. We have fitted the logistic regression model in (2) to these data with the addition of a covariate for maternal age (MA). Results are very similar to that shown above as the odds ratio relating TGFA to the risk for oral cleft was estimated at 1.00 ($= e^{0.0014}$) for children whose mothers do not smoke and as 1.09 ($= e^{0.0014+0.088}$) for children whose mothers do smoke. Thus, these data provide little evidence of interaction between TGFA and maternal smoking regarding risk for oral cleft.

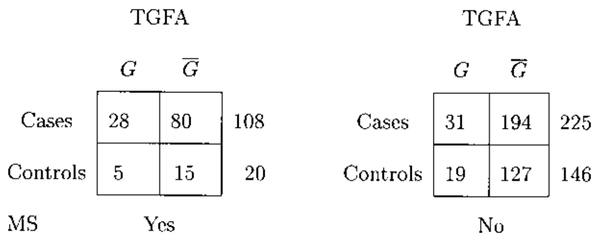
Finally, sample size calculations have been developed to detect gene–environment interactions in unmatched case-control studies.[7–9] More recently, Sturmer and Brenner[10] pointed out through simulations that considerable gain in power for testing interactions may result by matching (frequency or individual) on the environmental factor in the design stage, especially if the environmental factors have low prevalence in the population. Such a gain in statistical power may be offset by the extra difficulty in identifying matched controls for the very reason of low prevalence of environmental factors. Sturmer and Brenner[10] suggest the balance between power gain and extra costs to achieve matching must take into account the specific research questions and surrounding circumstances.

2.3 Searching for homogeneous subgroups

Under the framework of case-control designs, one can test for homogeneity among different subgroups of cases. Assuming that a hypothesized subtyping variable, such as early versus late onset of the disease, is available the following polytomous logistic regression model can be fit as

$$\log \frac{\Pr(Y = j|\mathrm{FH(S)})}{\Pr(Y = 0|\mathrm{FH(S)})} = \alpha_j + \beta_j \mathrm{FH(S)}, \quad j = 1, \ldots, C \tag{3}$$

where Y is a categorical response of $C + 1$ categories with $Y = 0$ for controls and $Y = j$ for cases of a particular subtype, $j = 1, \ldots, C$. For the onset age example, C would be 2 with $Y = 2(1)$ if the case diagnosed with late (early) onset. If the regression coefficients (β_j's) are significantly different, then more homogeneous subgroups can be identified and may be targeted for further investigations.



	TGFA		
	G	$\overline{G}$	
Cases	28	80	108
Controls	5	15	20
MS	Yes		

	TGFA		
	G	$\overline{G}$	
Cases	31	194	225
Controls	19	127	146
	No		

Figure 3

Returning to the oral cleft study, it is also of interest to examine if the two anatomical subtypes of oral cleft, cleft palate only (CP) and cleft lip with/without palate (CLP), are heterogeneous aetiologically. The 2 × 3 table in Figure 4 shows that the prevalence of carrying C2 allele is highest among cases with CP (21.8%), followed by CLP cases (15.3%) and 14.5% for the controls.

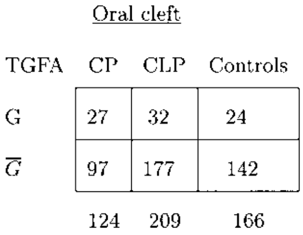
To formally answer this question, consider the following polytomous logistic regression model,

$$\log \frac{\Pr(Y = j|\text{TGFA, MS, MA})}{\Pr(Y = 0|\text{TGFA, MS, MA})} = \alpha_j + \beta_{j1}\text{TGFA} + \beta_{j2}\text{MS} + \beta_{j3}\text{TGFA} \star \text{MS} + \beta_{j4}\text{MA}$$

where $Y = 2(1)$ if the case diagnosed with CP(CLP) and $Y = 0$ for controls. Results are shown in Table 1. It appears that for children of non-smoking mothers, there is no association between TGFA and the risk of oral cleft for either types of oral clefts ($e^{\hat{\beta}_{11}} = 1.05$ and $e^{\hat{\beta}_{21}} = 0.98$ for CP and CLP, respectively) and hence very little evidence of heterogeneity in genotype effect between these two types of oral cleft. However, for children whose mothers smoke, there is a stronger positive association between TGFA and the risk of being born with CP ($e^{\hat{\beta}_{11}+\hat{\beta}_{13}} = 1.87$) compared to the association between TGFA and the risk of being born with CLP ($e^{\hat{\beta}_{21}+\hat{\beta}_{23}} = 0.74$). Although the difference between these two odds ratios is not statistically significant at the 0.05 level, it raises the possibility that these two subtypes of oral cleft may have different genetic aetiology with respect to TGFA, perhaps modified by maternal smoking.

2.4 Pros and cons of this design

Compared to the other design to be discussed later, this conventional case-control design has the advantage of being easily implemented and very familiar to researchers in epidemiology. For example, there is no direct interviews of relatives required and the information concerning the primary 'risk factor', family history of the disease, can be collected through cases/controls themselves (or their informants). On the other hand, the quality of the family history data may be questioned as discussed in Section 2.1. This in turn may lead to potentially erroneous conclusions concerning the presence and the degree of familial aggregation. While some simple alternatives such as creating a family history score have been suggested to alleviate these concerns, the fact remains that it is not obvious that data from a case-control design can always be



	Oral cleft		
TGFA	CP	CLP	Controls
G	27	32	24
$\overline{G}$	97	177	142
	124	209	166

Figure 4

Table 1 Polytomous logistic regression estimates (± s.e.) for the oral cleft study[6]

Variable	CP/control	CLP/control
Intercept	2.756	3.388
	(0.753)	(0.679)
TGFA	0.045	–0.025
	(0.406)	(0.580)
MS	0.821	1.071
	(0.370)	(0.329)
TGFA*	0.580	–0.279
MS	(0.746)	(0.714)
MA	–0.108	–0.112
	(0.024)	(0.022)

CP: cleft palate only; CLP: cleft lip with/without palate.

carried to the next steps of scientific inquiry (e.g. segregation and linkage analysis) should evidence of familial aggregation be substantiated. For these reasons, we consider in the next section an alternative design.

Finally, a few technical comments concerning the use of logistic regression models for all three issues addressed above. First, one reason this model is chosen is because the regression coefficients, with the exception of the intercept, are estimable under the case-control design. Secondly, this model is flexible in that it allows adjustment through regression for confounding variables, whether matched in the design stage or not. Thirdly, if individual matching is part of the case-control design, however, the conventional likelihood inference for either unmatched or frequency matching design[11] is known to fail and the alternative inferential procedure, known as the conditional logistic regression method, may be adopted.[5]

3 Family case-control design

3.1 Description

During the past decade, this design has drawn a good deal of attention among researchers in genetic epidemiology to address issues considered here; see, for example Claus *et al.*[12] and Mettlin *et al.*[13] for breast cancer, Pulver and Liang[14] for schizophrenia and more recently, Nestadt *et al.*[15] for obsessive compulsive disorders. Specifically, with the consent of cases and controls, their relatives are recruited and interviewed directly for detailed evaluations on their disease status, laboratory assessments and demographic and risk factor information relevant to the disease. For this design, the phrase ‘case (control) proband’ has been coined for cases (controls) as they represent individuals through which the family is ascertained. Just as in Section 2, the data can be summarized in a 2×2 (Figure 5).

The primary response in this design is the risk among relatives, known as the familial risk. Thus, familial aggregation of the disease may be claimed if the risk of disease among case relatives is substantially higher than that among control relatives.

	Disease status	
	Affected	Unaffected
Case relative		
Control relative		

Figure 5

It is important to point out that a primary difference between the conventional case control design and this family case-control design lies in the sampling unit and on the quantity to be compared with. For the former, the unit, either from the sampling or analytical viewpoint, is an individual (case and control) and comparison is made between cases' characteristics and controls' characteristics, e.g. family history or a genetic marker. For the latter, however, the unit is family in which characteristics of the individual relatives, and disease status within families is compared between case families and control families. This distinction has profound implications on validity of statistical inferences and on practical implementation as addressed in the rest of this section.

3.2 To match or not to match

Just as in conventional case-control studies, one has the choice of matching each case with a control or not. However, one may argue that the matching criteria for family case-control designs should be subject to some modification. Recall a major principle behind matching is to assure, to the extent possible, that 'units' to be compared are indeed comparable. Given that the comparison is made between case relatives and control relatives, finding a control comparable to a case may not serve this purpose. Thus for family case-control designs, matching at the design stage may be warranted if the primary confounding variables are themselves highly familial. This would increase the likelihood that case relatives and matched control relatives are comparable to each other regarding the matching variables. Confounding variables which are not necessarily familial, e.g. gender, can be easily adjusted through regression.

Whether matching or not, statistical methods that have been fully developed for the conventional case-control design, including the conditional logistic regression analysis for matched designs, are not adequate here. This is because, as pointed out toward the end of Section 3.1, the unit is a family and response variables (i.e. affected status of relatives) are unlikely to be statistically independent of each other. This additional analytical complication, namely, within-family correlation in risk, can be dealt with accordingly. For unmatched designs, one may, for example, consider the use of the generalized estimating equation (GEE) method which was specifically developed to take into account this complication.[16,17] As for the matched design, the method developed by Liang[18] could be used to address the issues considered here while accounting for correlations of among related individuals. These two methods have been successfully applied to some genetic studies that address the three issues as illustrated in the rest of this section. For a more detailed discussion on the utility of this design and two analytical methods mentioned above for other scientific questions of interest, and the issue of sample size calculations, see Liang and Pulver.[19] Finally, for age of onset outcomes, methods[20–22] for detecting familial aggregation have also been developed to take into account the within-family correlation complication.

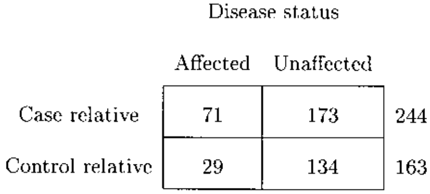
3.3 Detecting familial aggregation and testing gene–environment interactions

To see how the family case-control design may be used to more formally address the issue of familial aggregation, let $Y = (Y_1, \ldots, Y_n)$ be the disease status of n relatives from a family ascertained through a proband (either a case or control). Consider for each relative $j, j = 1, \ldots, n$,

$$\text{logit } \Pr(Y_j = 1) = \alpha + \beta^t x_j + \gamma z \tag{4}$$

where $z = 1(0)$ if the proband from which the jth relative was sampled is a case (control) proband and x_j the covariates specific to the jth relative or to the proband. The key parameter of interest is obviously γ which, in the absence of any x, corresponds to the log odds ratio from the 2×2 table displayed in Section 3.1. In general, this parameter characterizes the overall difference in familial risk, i.e. $\Pr(Y = 1)$, between case families and control families allowing for differential risk among individuals with different observed risk factor values, x_j. However, it is known that ignoring any correlation in Y among relative individuals would lead to incorrect estimates of the variance about these regression estimators in (4).[17] This concern as pointed out in Section 3.2 can be alleviated by adopting the GEE method for unmatched designs and the extended Mantel–Haenszel method by Liang[18] for matched designs.

Returning to the COPD study where first-degree relatives of case and control probands were directly examined for pulmonary function. Table 2 shows frequencies of affected relatives by case/control status and family size. Thus, for example, among 27 case families with three first degree relatives sampled, nine of them have one (out of three) relatives affected. Data from Table 2 can be summarized into a single 2×2 table as shown in Figure 6. Here 71 out of 244 case relatives were diagnosed with impaired pulmonary function (29%), whereas 29 out of 163 control relatives experienced the same condition (18%). This leads to an estimated log odds ratio of 0.64 (with s.e. $= 0.25$), suggesting that the familial risk among case families is twice ($\equiv e^{0.64}$) as high as that of control families. This simple approach may be criticized on the grounds that the s.e. estimate is likely to be too conservative due to its failure to



	Disease status		
	Affected	Unaffected	
Case relative	71	173	244
Control relative	29	134	163

Figure 6

Table 2 Frequencies of affected relatives by case/control status and number of relatives per proband from the COPD Study[3]

	Number of impaired pulmonary functions						
Number of relatives	0	1	2	3	4	6	Total
Case family							
1	25	12	–	–	–	–	37
2	13	12	1	–	–	–	26
3	12	9	4	2	–	–	27
4	3	4	1	0	0	–	8
5	1	1	0	0	0	–	2
6	0	0	0	1	1	1	3
7	1	0	1	0	0	0	2
Control family							
1	32	8	–	–	–	–	40
2	16	5	1	–	–	–	22
3	4	1	0	0	–	–	5
4	2	0	1	1	0	–	4
5	2	2	0	0	0	–	4
6	0	3	0	0	0	0	3
7	0	0	0	0	0	0	0
10	0	0	0	1	0	0	1

account for within-family correlation in Y, and that important risk factors such as smoking status were not properly considered. To alleviate these concerns, we employed the logistic regression model in (4) with the GEE method to the same data and results are presented in Table 3. Compared with the analysis presented above, results in Model I suggest that the correct s.e. of $\hat{\gamma}$ is indeed larger (s.e. $= 0.28$), although the discrepancy (0.25 versus 0.28) is not substantial in this example. Results from Model II are interesting in that after adjusting for some key individual risk factors including smoking, age, etc. stronger evidence of familial aggregation is revealed. Specifically, the familial risk in case families is now estimated to be 2.23 ($= e^{0.80}$) times as high as that in control families.

To test for gene–environment interactions under this framework, one can simply add in (4) interaction terms between some of the xs and z. Model III in Table 3 examines the interaction between z and smoking status. While not statistically significant, results suggest the effect of smoking is considerably higher among case families ($\widehat{\mathrm{OR}} = e^{0.91+0.14} = 2.86$) than that among control families ($\widehat{\mathrm{OR}} = e^{0.14} = 1.15$).

3.4 Searching for homogeneous subgroups

Family case-control designs are particularly useful for identifying subgroups if the candidate variables for subtyping are clinical characters associated with the occurrence of the disease. Examples include age at onset (early versus late) of breast cancer[12,13] and of schizophrenia[14] and presence or absence of 'flow lesion' defects in patients diagnosed congenital cardiovascular malformation (CCVM).[23] In this situation, we can expand the logistic regression model in (4) by allowing multiple zs, i.e.

$$\operatorname{logit} \Pr(Y_j = 1) = \alpha + \beta^t x_j + \gamma_1 z_1 + \ldots + \gamma_C z_C \qquad (5)$$

Table 3 Logistic regression estimates (± s.e.) using the GEE method to assess familial aggregation in impaired pulmonary function and interaction with smoking status from the COPD Study[3]

		Models	
Variable	I	II	III
Intercept	–1.54	–1.57	–1.09
	(0.23)	(0.34)	(0.41)
Sex		–0.79	–0.73
(1: Male)		(0.28)	(0.27)
Race	–0.55	–0.57	
(1: White)		(0.29)	(0.30)
Smoking		0.75	0.14
(1: Yes)		(0.30)	(0.47)
Age		0.061	0.058
(in years)		(0.014)	(0.014)
Parent		–0.25	–0.32
(1: Parent)		(0.37)	(0.35)
Proband	0.64	0.80	0.26
(1: Case)	(0.28)	(0.31)	(0.45)
Proband*		0.91	
Smoking			(0.56)

where, with the control probands as the reference group, C represents the number of subgroups among all case probands. For the CCVM example, C would be two with

$$z_1 = \begin{cases} 1 & \text{if the case proband has a flow lesion} \\ 0 & \text{o.w.} \end{cases}$$

$$z_2 = \begin{cases} 1 & \text{if the case proband does not have a flow lesion} \\ 0 & \text{o.w.} \end{cases} \tag{6}$$

A variable characterized by the zs would be useful for genetic subtyping if some of the γs are different from each other and presumably, subgroups having larger familial risks would be the ones targeted first for further investigation.

We now illustrate the use of model (5) and the statistical method developed by Liang[18] for a matched family case-control study on CCVM.[23] Here each of 570 cases who have one or more full sibs (363 with flow lesion defects and 207 without) was matched with a control born within the same 30-day period and also having at least one full sib. Among 1963 case relatives (1140 parents and 823 siblings), 41 were diagnosed with CCVM; whereas only 10 out of 1946 control relatives (1140 parents and 806 siblings) were affected. While strong evidence of familial aggregation was found overall (2.1% versus 0.05%), this *ad hoc* approach ignores the matching aspect of the design and individual risk factors such as sex were not considered. After adjusting for race, gender and relationship to probands (parents versus siblings) and adopting the method by Liang,[18] Maestri *et al.*[23] found that γ in (4) is estimated as 1.405 (s.e $= 0.424$). This implies that the familial risk among case families is 4.14 ($= e^{1.405}$) times as high as among relatives of controls, again showing a strong evidence of familial aggregation. To further test the hypothesis that the familial risks for flow lesion and non-flow lesion case probands are different, Maestri *et al.*[23] considered the model specified by (5) and (6) and found that $\hat{\gamma}_1 = 1.698$ (s.e. $= 0.498$) and $\hat{\gamma}_2 = -0.330$ (s.e. $= 0.765$). This suggests that the excessive familial risk among relatives of cases seen above is mainly attributed to those with flow lesion and there is strong evidence of aetiologic heterogeneity between two types of CCVM as

$$\frac{(\hat{\gamma}_1 - \hat{\gamma}_2)^2}{\mathrm{Var}(\hat{\gamma}_1 - \hat{\gamma}_2)} = 4.94$$

was statistically significant at the 0.05 level.

3.5 Pros and cons of this design

Compared to the conventional case-control design, this approach requires the direct recruitment of probands' relatives and direct detailed evaluations of these relatives regarding disease status, relevant demographic information, and perhaps laboratory assessment. This amounts to a much greater degree of commitment of cost and human resources. On the other hand, this additional burden may be offset by the following advantages. First, through direct assessment of affected status of relatives, some of the concerns such as recall bias, misclassification of family history associated with the use of conventional case-control designs can be avoided. Secondly, this design provides a direct estimate of recurrence risk (probability for being affected given someone in the family is affected), which can be useful for genetic counselling. For the COPD study, this empiric recurrence risk is estimated as 0.29 ($= 71/244$); see Section 3.3. Thirdly, assuming controls are representative of the general population, e^{β} in (4) provides an approximation to λ, known as the risk ratio[24], which is the ratio of recurrence risk to the risk of general population. This quantity plays a crucial role in sample size calculation for the number of affected sib pairs needed in linkage analysis.[24] Fourthly, with relatives of probands recruited, one is in the position to conduct formal segregation analysis to identify a genetic mechanism underlying the disease. Furthermore, it is common that multiplex families, i.e. families with two or more affected, can be extracted from families ascertained through a case to carry out linkage analysis for the purpose of localizing susceptibility genes. This rather smooth transition from the tasks of detecting familial aggregation, searching for homogeneous subgroups, etc. to the more intensive tasks of segregation and linkage analyses is not shared by conventional case-control designs. Finally, family case-control design is likely to be more efficient statistically for the same task than the conventional case-control design since the primary measure of familial aggregation, family history (FH), for conventional case-control studies is subject to the limitations discussed in Section 2. To illustrate, the 2 × 2 table in Figure 7 (on the top) shows proportions of affected relatives among 44 (241) male probands diagnosed with schizophrenia before (after) reaching 16 years of age.[14] Adjusting for each relative's gender and age, GEE analysis showed the familial risk among relatives of early onset probands is 2.5 ($\hat{\gamma} = 0.91$ with s.e. $= 0.44$) times as

Case relatives	Schizophrenia		
of probands with	Affected	Unaffected	
early onset	8	191	199
late onset	20	1161	1181

Case	Family history		
probands	Positive	Negative	
early onset	7	37	44
late onset	18	223	241

Figure 7

high as that among relatives of late onset probands. This suggests that patients diagnosed with schizophrenia at early ages may have different genetic aetiology compared those diagnosed at later ages. On the other hand, if one adopts the approach by simply comparing positive family history of schizophrenia between cases with either early and late onset ages (see the 2 × 2 table on the bottom), one leads to an estimate of log odds ratio 0.85 with s.e. 0.48, yielding no statistically significant evidence for subtyping in onset age.

4 Measuring familial correlations

An important advantage of the family case-control design is its ability to measure familial correlations in terms of both their magnitude and patterns. It is intuitive that the larger the magnitude of familial correlation, the stronger the evidence for some genetic basis in determining disease. This information about the magnitude of familial risk has a significant bearing on the degree of statistical power available to locate susceptibility genes. Furthermore, examining the patterns of familial aggregation closely allows investigators to further tease apart the roles played by genetic and environmental factors. For example, with trait data from first degree relatives of probands, higher correlations in risk among siblings compared to parents could suggest a dominance effect of unobserved gene(s). On the other hand, higher correlations between mothers and offspring than fathers and offspring provides a clue that the genetic mechanism is not a simple autosomal mechanism. Thus, while examining familial correlations is at best viewed as an exploratory data exercise where assumptions about genetic mechanisms are minimal, it may provide invaluable information and clues concerning the underlying genetic mechanism and its possible interaction with environmental factors.

In this section, we briefly review measures for familial aggregation from the literature. Depending on the type of phenotypic traits dealt with, e.g. quantitative, qualitative or survival data subject to censoring (such as onset age) different measures of familial correlation may be employed. Our focus is on how these measures may be modelled statistically through regression techniques and be estimated should the family case-control design be adopted.

4.1 Quantitative traits

For quantitative traits observed on pairs of relatives (e.g. twins) such as cholesterol level or birthweight, Y_1 and Y_2 say, the most commonly used measure of familial correlation is Pearson's product moment correlation coefficient. This intra-class correlation is defined as

$$\rho = \mathrm{Cov}(Y_1, Y_2)/(\mathrm{Var}(Y_1)\mathrm{Var}(Y_2))^{1/2} \tag{7}$$

For designs involving twins only, this correlation also provides a direct estimate of heritability, the proportion of total variation due to genetic variance due to one or more autosomal loci. For sets of three or more relatives such as sibships, an intra-class correlation can be used which is defined as the ratio of variance among sibships over the sum of the within sibship variance and the among sibship variance. This can be viewed as the average intra-class correlation over all possible pairs of sibs. Clearly, the higher the ρ, the stronger the evidence of familial resemblance. While such familial resemblance may have no genetic basis, careful design and modelling of ρ should provide clues on the roles genetic factors may play.

For a family of size n, we consider a general model for ρ_{jk}, the correlation coefficient for a quantitative trait Y, for the jth and kth relatives, $j < k = 1, \ldots, n$

$$\mathrm{logit}\left(\frac{1+\rho_{jk}}{2}\right) = \alpha_0 + \alpha^t z_{jk} \tag{8}$$

where z_{jk} is a set of q covariates which could be specific to the (j, k) pair, specific to the family or some combination of both.[25] The transformation on the left-hand side of (8), $\mathrm{logit}\{(1+\rho)/2\}$, ensures that this measure ranges over the whole real numbers. As an illustration, we now consider several examples regarding the utility of (8) for testing different hypotheses.

Example 1

For designs including nuclear families consisting of parents and offspring, several different correlations should be considered

$$\rho_{jk} = \rho_{\mathrm{SS}}, \rho_{\mathrm{PS}} \text{ or } \rho_{\mathrm{PP}}$$

depending on whether the (j, k) relatives pair are siblings, a parent and an offspring or two parents, respectively. Of particular interest is the comparison between ρ_{SS} and ρ_{PP}. Assuming all relatives share a similar environment, a significantly higher ρ_{SS} than ρ_{PP} would strengthen the argument for genetic determination of the trait. On the other hand, one can test the hypothesis of, for example, maternal transmission mechanism by contrasting ρ_{MS} with ρ_{FS}, where $\rho_{\mathrm{MS}}(\rho_{\mathrm{FS}})$ is the pairwise ρ between mother (father) and each offspring.

Example 2

For designs involving siblings only, one may readily test the hypothesis that the constant within-family correlation is associated with an observable family-specific covariates, for example race. For example, let $z_{jk} = 1(0)$ if whites (blacks) and fit separate correlation coefficients for blacks and whites. This approach also allows investigators to identify variables that may reflect heterogeneity among subgroups of families.

Example 3

While a constant sibling correlation does provide a measure of heritability, one can further examine risk factors that may influence this sibship correlation using (8). For traits such as birthweight, one can, for example, model ρ_{jk} as a function of the time between any two pregnancies of sibs j and k.

To make inference on the ρs, one can take the likelihood approach by assuming $Y = (Y_1, \ldots, Y_n)$ from a family size n follows a multivariate normal distribution with the covariance matrix consistent with the ρs that are specified. An alternative is to take the estimating function approach[17] which may be viewed as a multivariate analogue of the quasi-likelihood method of Wedderburn.[26] In either case, it is imperative that important risk factors for the Ys be properly considered. This can be achieved by modelling

$$\mathrm{E}(Y_j|x_j) = x_j^t\beta,$$

where x_j is the observed covariates for the jth relative, $j = 1, \ldots, n$. For applications of the estimating function approach and use of (8) to measure familial correlations of birth weight, see Beaty *et al*.[25]

4.2 Qualitative traits

For qualitative phenotypic traits such as disease status (affected or unaffected), the correlation coefficient is a less desirable measure of familial correlation because the range of ρ is constrained, sometimes severely, by the prevalence of the trait.[27] For example, if the disease prevalence among relatives is 0.3, the range of ρ is restricted to $(-0.33, 0.76)$. While many measures of association for categorical variables have been proposed,[28] we recommend the use of the odds ratio as a measure of familial correlation[29] as (i) it is familiar to researchers in public health, (ii) there is no constraint on this measure and (iii) it can be extended easily to ordinal variables as well.[30] Formally, an odds ratio between two qualitative traits is defined as

$$\mathrm{OR}_{jk} = \frac{\Pr(Y_j = 1 = Y_k)\Pr(Y_j = 0 = Y_k)}{\Pr(Y_j = 1, Y_k = 0)\Pr(Y_j = 0, Y_k = 1)}$$

An OR greater than one suggests that positive familial association between risk to relatives and the larger this OR, the stronger this familial association. As in (8), one can model OR_{jk} by considering

$$\log \mathrm{OR}_{jk} = \alpha_0 + \alpha^t z_{jk} \tag{9}$$

This measure of familial correlation has recently been applied to a genetic epidemiologic study of liver cancer by Dr Shen[29] in which the GEE method developed by Liang *et al.*[27] was employed for making inferences about the αs. Again, it is important that relevant risk factors for the qualitative trait be properly considered in the model. This could be achieved by considering, for example, the following logistic regression model for each subject,

$$\mathrm{logit}\Pr(Y_j = 1|x_j) = x_j^t\beta$$

As shown in Liang and Beaty,[29] such adjustments for individual risk factors may have a profound impact on inference for magnitude of familial correlations as characterized by the αs in (9).

4.3 Onset ages

For disease or disorders with low incidence rates and age-specific risk, it is likely that many subjects at the time of data collection are unaffected, a phenomenon known as censoring. Treating these subjects as unaffected could result in substantial loss of statistical power especially if onset age spans a wide range. A more sensible approach is to use as the trait the age at onset to acknowledge statistically that some of the data may be censored. This is known as survival analysis. For survival responses, several measures of familial correlations have been suggested. One is to use the correlation coefficient for $M(Y_j)$ and $M(Y_k)$, where M is a function of Y, the onset age. These include as a special case ρ itself,[31] i.e. $M(Y) = Y$ and the cumulative hazards function of Y.[32] We consider an alternative by appealing to the concept of hazard functions directly. Specifically, denote $\lambda(t)$ a proper hazards function at time t, we consider the ratio[33]

$$\theta = \frac{\lambda(Y_j = t_j|Y_k = t_k)}{\lambda(Y_j = t_j|Y_k > t_k)} = \frac{\lambda(Y_k = t_k|Y_j = t_j)}{\lambda(Y_k = t_k|Y_j > t_j)} \tag{10}$$

No association between Y_j and Y_k is declared if $\theta = 1$ and a positive association in onset ages may be claimed if $\theta > 1$. A probability function for $Y = (Y_1, \ldots, Y_n)$ that is consistent with the assumption of constant θ for all possible $\binom{n}{2}$ pairs does exist and can be expressed as[33,22]

$$\Pr(Y_1 > t_1, \ldots, Y_n > t_n) = \left(\sum_{j=1}^{n} S_j^{1-\theta}(t_j) - (n-1) \right)^{-1/(\theta-1)} \tag{11}$$

where $S_j(t_j)$ is the marginal survival function for $Y_j, j = 1, \ldots, n$. A pseudo-likelihood procedure[34] has been developed to estimate θ when the S_js are either modelled nonparametrically,[35,36] semi-parametrically[37] or parametrically.[38] Furthermore, Bandeen-Roche and Liang[37] have extended (11) to more general situations where first degree relatives are sampled (e.g. parents and offspring) and thus more than one θ parameter may be required. Finally, Pulver and Liang[14] have applied (11) to a genetic epidemiologic study of schizophrenia[39] to test the hypothesis that onset age (e.g. early versus late) may be useful for genetic subtyping.

4.4 Ascertainment issue

A common phenomenon in genetic epidemiologic research is that the sampled families are not representative of the targeted population as they are ascertained through probands with known phenotypic values. It is well known in the literature of segregation analysis that statistical inference without proper ascertainment corrections would lead to biased estimations of key parameters such as segregation ratios, penetrances, etc. This would also be the case for estimating familial correlations should, for example, the family case-control design be adopted. One simple remedy,[40] which is commonly adopted, is to condition on the observed phenotypic values of the probands (case or control). Specifically, let $\Pr(Y_1, \ldots, Y_n; \theta, \phi)$ be the joint probability (density) function for a sampled family of size n ascertained through Y_1, the phenotype of the proband. Here θ represent familial correlation parameters and ϕ the nuisance parameters (e.g. the underlying hazard functions in (11)) that are needed to fully specify the joint distribution for $(Y_1, \ldots, Y_n)$. Then one may eliminate (or at least reduce) the ascertainment bias by basing the inference on

$$\Pr(Y_2 = y_2, \ldots, Y_n = y_n | Y_1 = y_1; \theta, \phi) \tag{12}$$

the conditional distribution for the non-proband members, given the phenotypic value of the proband. It is important to point out that through this conditioning, some of the nuisance parameters in ϕ may become non-identifiable. This restriction, however, will not limit one's ability to estimate θ using (12). A well-known analogue is in the conventional case-control design in which all of the logistic regression coefficients are estimable with the exception of the intercept.

In the absence of full knowledge of the joint distribution for $(Y_1, \ldots, Y_n)$, one approach is to express, for each $j = 2, \ldots, n$, the $\mathrm{E}(Y_j|Y_1)$ in terms of θ which in turn can be estimated through, for example, the GEE method by equating Y_j to $\mathrm{E}(Y_j|Y_1)$ for all $j = 2, \ldots, n$.

5 Discussion

In this paper, we review statistical designs useful to address a central question in genetic epidemiology, namely, finding evidence for familial aggregation or clustering in disease risk. While not exhaustive, the two designs considered here are commonly used. We discuss the pros and cons of each design and its associated analytical complications. If a positive finding is reported, a natural question is to what extent is the familial aggregation in disease accounted for by familial aggregation of some environmental factor(s)? For the COPD study, one may wonder that the important risk factor, smoking habit, may itself be clustered in families, and ask could this explain the clustering in impaired pulmonary function. The work by Khoury *et al.*[41] suggests, however, that unless these environmental factors are associated with extreme relative risks and demonstrate extreme patterns of familial clustering, it is very unlikely that the risk factor alone would account for familial aggregation of disease. For example, Khoury *et al.*[41] show that even if the relative risk relating the environmental factor to the disease is as high as 10 and the degree of familial correlation for this risk factor is as high as 0.5, the maximum risk ratio of comparing the risk of disease for case siblings with that for control siblings is 2.01. Thus in the absence of extreme patterns, a well designed study leading to a high estimate of risk ratio or correlation would suggest that such a finding is unlikely to be explained by familial clustering in risk factors.

In addition to the central question of detecting familial aggregation, we discussed several closely related issues and showed how the same design can be utilized to address those issues as well. These issues include testing gene–environment interactions, searching homogeneous genetic subgroups and measuring familial correlations. An interesting feature of the analytical strategies considered here is that they are exploratory in nature and require no prior specification of any genetic mechanism. Findings from these issues, however are potentially extremely useful for subsequent tasks including segregation and linkage analyses. For example, the work by Claus *et al.*[12] and Mettlin *et al.*[13] showing familial risk in breast cancer is higher among probands with early onset has led to the important linkage finding in chromosome 17.[42]

Acknowledgement

This work is supported in part by Grant GM49909 of the National Institutes of Health.

References

1 Andrieu N, Goldstein AM. Epidemiologic and genetic approaches in the study of gene-environment interaction: an overview of available methods. *Epidemiology Review* 1998; **20**: 137–147.

2 Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* 1959; **22**: 719–48.

3 Cohen BH. Chronic obstructive pulmonary disease: A challenge in genetic epidemiology. *American Journal of Epidemiology* 1980; **112**: 274–88.

4 Khoury MJ, Beaty TH, Cohen BH. *Fundamentals of genetic epidemiology*. London: Oxford University Press, 1993.

5 Breslow NE, Day NE. *Statistical methods in cancer research*, Vol. 1. Lyon: IARC Scientific Publication No. 32, 1980.

6 Beaty TH, Maestri NE, Hetmanski JB *et al.* Testing for gene–environment interaction among oral cleft cases born in Maryland 1992–1996. *Journal of Palate-Craniofacial* 1997; **34**: 447–54.

7 Hwang S-J, Beaty TH, Liang K-Y, Coresh J, Khoury MJ. Minimum sample size estimation to detect gene–environment interaction in case-control designs. *American Journal of Epidemiology* 1994; **140**: 1029–37.

8 Foppa I, Spiegelman D. Power and sample size calculations for case-control studies of gene–environment interactions with a polytomous exposure variable. *American Journal of Epidemiology* 1997; **146**: 596–604.

9 Garcia-Closas M, Lubin JH. Power and sample size calculations in case-control studies of gene–environment interactions: comments on different approaches. *American Journal of Epidemiology* 1999; **149**: 689–92.

10 Sturmer T, Brenner H. Potential gain in efficiency and power to detect gene–environment interactions by matching in case-control studies. *Genetic Epidemiology* 2000; **18**: 63–80.

11 Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika* 1979; **66**: 403–11.

12 Claus EB, Risch NJ, Thompson WD. Age at onset as an indicator of familial risk of breast cancer. *American Journal of Epidemiology* 1991; **131**: 961–72.

13 Mettlin C, Croghan I, Natarajam N, Lane W. The association of age and familial risk in a case-control study of breast cancer. *American Journal of Epidemiology* 1991; **131**: 973–83.

14 Pulver AE, Liang K-Y. Estimating effects of probands' characteristics on familial risk: II. The association between age at onset and familial risk in the Maryland Schizophrenia Sample. *Genetic Epidemiology* 1991; **8**: 339–50.

15 Nestadt G, Samuels J, Bienvenu OJ *et al*. A family study of obsessive compulsive disorder. *Archives of General Psychiatry* 2000; **57**: 358–63.

16 Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.

17 Liang K-Y, Zeger SL. Regression analysis for correlated data. *Annual Review of Public Health* 1993; **14**: 43–68.

18 Liang K-Y. Extended Mental-Haenszel estimating procedures for multivariate logistic regression models. *Biometrics* 1987; **43**: 289–99.

19 Liang K-Y, Pulver AE. Analysis of case-control/family sampling design in psychiatric research. *Genetic Epidemiology* 1996; **13**: 253–70.

20 Hsu L, Zhao L. Assessing familial aggregation of age of onset, by using estimating equations, with application to breast cancer. *American Journal of Human Genetics* 1996; **58**: 1057–71.

21 Li H, Yang P, Schwartz AG. Analysis of age of onset data from case-control family studies. *Biometrics* 1998; **54**: 1030–39.

22 Liang KY. Estimating effects of probands' characteristics on familial risk: I. Adjustment for censoring and correlated ages at onset. *Genetic Epidemiology* 1991; **8**: 329–38.

23 Maestri NE, Beaty TH, Liang K-Y, Boughman JA, Ferencz C. Assessing familial aggregation of congenital cardiovascular malformations in case-control studies. *Genetic Epidemiology* 1988; **5**: 343–54.

24 Risch N. Linkage strategies for genetically complex traits. I. Multilocus models. *American Journal of Human Genetics* 1990; **46**: 222–28.

25 Beaty TH, Skjaerven R, Breazeale DR, Liang K-Y. Analyzing sibship correlations in birth weight using large sibships from Norway. *Genetic Epidemiology* 1997; **14**: 423–33.

26 Wedderburn RWM. Quasi-likelihood function, generalized linear models and the Gaussian method. *Biometrika* 1974; **61**: 439–47.

27 Liang K-Y, Zeger SL, Qaqish BF. Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society Series B* 1992; **54**: 3–40.

28 Goodman LA, Kruskal WH. *Measure of association for cross-classifications*. New York: Springer, 1979.

29 Liang K-Y, Beaty TH. Measuring familial aggregation by using odds-ratio regression models. *Genetic Epidemiology* 1991; **8**: 361–70.

30 Heagerty PJ, Zeger SL. Marginal regression models for clustered ordinal measurements. *Journal of the American Statistical Association* 1996; **91**: 1024–36.

31 MacLean CJ, Neale MC, Meyer JM, Kendler KS. Estimating familial effects on age at onset and liability to schizophrenia II. Adjustment for censored data. *Genetic Epidemiology* 1990; **7**: 419–26.

32 Prentice RL, Cai J. Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika* 1992; **79**: 495–512.

33 Clayton DG. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* 1978; **65**: 141–51.

34 Gong G, Samaniego FJ. Pseudo maximum likelihood estimation: Theory and applications. *Annals of Statistics* 1981; **89**: 861–69.

35 Genest C, Ghoudi K, Rivest L-P. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika* 1995; **82**: 543–52.

36 Shih JH, Louis TA. inferences on the association parameter in Copula models for bivariate survival data. *Biometrics* 1995; **51**: 1384–99.

37 Bandeen-Roche KJ, Liang K-Y. Modeling failure-time association in data with multiple levels of clustering. *Biometrika* 1996; **83**: 29–39.

38 Huster WJ, Brookmeyer R, Self SG. Modelling paired survival data with covariates. *Biometrics* 1989; **45**: 145–56.

39 Pulver AE, Bale SJ. Availability of schizophrenic patients and their families for genetic linkage studies: findings from the Maryland Epidemiology Sample. *Genetic Epidemiology* 1989; **6**: 671–80.

40 Ewens WJ, Shute NE. A resolution of the ascertainment sampling problem. I. Theory. *Theoretical Population Biology* 1986; **30**: 388–412.

41 Khoury MJ, Beaty TH, Liang K-Y. Can familial aggregation of disease be explained by familial aggregation of environmental risk factors? *American Journal of Epidemiology* 1988; **127**: 674–83.

42 Hall JM, Lee MK, Newman B, *et al*. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 1990; **250**: 1684–89.