12-1-2005

THE ROLE OF AN EXPLICIT CAUSAL FRAMEWORK IN AFFECTED SIB PAIR DESIGNS WITH COVARIATES

Constantine E. Frangakis
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, cfrangak@jhsph.edu

Fan Li
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, fli@jhsph.edu

Betty Q. Doan
Institute of Genetic Medicine, The Johns Hopkins University, bdoan@jhmi.edu

The role of an explicit causal framework in affected sib pair designs with covariates

Constantine E. Frangakis[1], Fan Li[1], and Betty Q. Doan[2]

[1]Department of Biostatistics, [2] Institute of Genetic Medicine

The Johns Hopkins University

615 N. Wolfe St., Baltimore, MD 21205, USA.

emails: cfrangak@jhsph.edu, fli@jhsph.edu, bdoan@jhmi.edu

Research Report FLD-11-05

December 1, 2005

SUMMARY. The affected sib/relative pair (ASP/ARP) design is often used with covariates to find genes that can cause a disease in pathways other than through those covariates. However, such "covariates" can themselves have genetic determinants, and the validity of existing methods has so far only been argued under implicit assumptions. We propose an explicit causal formulation of the problem using potential outcomes and principal stratification. The general role of this formulation is to identify and separate the meaning of the different assumptions that can provide valid causal inference in linkage analysis. This separation helps to (a) develop better methods under explicit assumptions, and (b) show the different ways in which these assumptions can fail, which is necessary for developing further specific designs to test these assumptions and confirm or improve the inference. Using this formulation in the specific problem above, we show that, when the "covariate" (e.g., addiction to smoking) also has genetic determinants, then existing methods, including those previously thought as valid, can declare linkage between the disease and marker loci even when no such linkage exists. We also introduce design strategies to address the problem.

KEY WORDS: Affected sib pairs; Causal effects; Genetic linkage; Partially controlled studies; Potential outcomes; Principal stratification.

1. Introduction

In order to identify genes that are responsible for a trait, researchers have long used the affected sibling (or relative) pairs design (ASP; Penrose, 1935; Risch, 1990a, and references therein). The main idea in these designs is that individuals with similar trait characteristics, as are affected siblings, should also have similar genetic characteristics in marker loci close to trait genes. A useful measure of genetic similarity at a marker is the number of alleles shared identically by descent (IBD), as its distribution is known under the null hypothesis of no linkage between the marker and the trait gene. It is well known, however, that such an approach alone, without using other information, can have low power to detect linkage (e.g., Risch, 1990b). Moreover, in order to understand the pathways through which the candidate genes act on the trait, methods need to use also other variables in the design and analyses (e.g., Olson, 1999).

Such methods have more recently been proposed for using information on covariates, that is here, factors that have been shown to explain variation in the main trait (Olson, 1999; Devlin et al., 2002). These methods essentially stratify on covariates and compare the IBD sharing to the null IBD distribution with no stratification. Such methods are now implemented by popular software (e.g. SAGE, 2003). However, there have been no clearly stated principles for why the IBD sharing under no linkage to the trait should (or should not) have the same distribution before and after stratification on the covariates. Such principles are important since stratification can capture linkage to the covariate, as opposed to the trait. Although this possible confounding has been noted as a possibility (e.g., Greenwood and Bull, 1999; Schaid, 2003), it is currently believed that a simple transformation of the covariates makes the analyses immune to this problem (see Schaid et al. 2003, p. 90, for the sum, and our Sections 3-4). For this or other reasons, little attention has been given to the validity of methods when covariates can have genetic determinants.

We show the need and we propose an explicit causal framework for methods to estimate the locus and the degree to which a gene directly affects the main trait, when using covariates with genetic determinants. We propose this causal framework in the context of a simple setting of affected sib pairs using potential outcomes and principal stratification. The general advantage of this framework is its ability to explicitly separate the different assumptions needed for inference, and thereby point to the different research hypotheses by which such assumptions can fail and be improved. In the specific ASP problem, we make explicit a set of assumptions that are more general than those usually made implicitly, and we show that as the studied markers are close to those determinants, the similarity created from the stratification on such covariates induces increased IBD sharing even under no physical linkage of the studied marker to the trait. We show that, as a consequence, standard methods for using covariates, even those currently believed immune, can declare linkage between the disease and marker loci even when no such linkage exists. We then indicate design methodology for addressing better the above goals.

2. Role of causal formulation in linkage with covariates

In order to make clear the meaning of different assumptions, our goal, and our methodological arguments, we introduce here a structure at the level of potential outcome for the trait and principal stratification of the covariate. We focus first on possible children of one mating pair of parents. We call "unit i" the product of a possible mating of the parents that includes the description (formation) of all the resulting DNA *except* the DNA at three particular loci, which we denote by l_G, l_Z and l_M. A zygote or offspring, can then be thought of as the result of attaching in unit i a possible genotype g, z, m at each of those loci (see Fig. 1). It is important to note that a "unit" is more basic than and can result in different zygotes, and hence offspring, depending on the result of meiosis and fertilization at those remaining three loci.

If a unit i is attached to genotypes g, z, m (by meiosis and fertilization), then for the resulting offspring we let: $Y_i(g, m, z)$ be the potential outcome (Rubin, 1978) that will be observed for the main trait of interest (1 for diseased, 0 otherwise), for example, an indicator for glaucoma by a certain age. We also let $X_i(g, z, m)$ be a potential covariate value, for example an indicator of alcohol consumption at that (or earlier) age.

In what follows, the framework of potential outcomes is, with respect to our goals, new in linkage analysis, and so it is important to briefly outline its purpose. Potential outcomes allow consideration of two possibly related but different processes: (I) the causal effect of a genotype on the potential outcome of a particular unit; and (II) the possible dependence between units' potential outcomes and the genotypes the units receive (the "assignment mechanism", Rubin, 1974, 1978). This separation is important because the main scientific interest here is in (I), whereas the observed data are a result of mixing (I) with (II). Therefore below, we first discuss a set of explicit assumptions in this framework, and state our goal and the induced assumptions on the observed data. Under those assumptions we show that the standard methods generally fail, and that better methods exist to address the goal. In future work we will subject those assumptions to scrutiny, examining when they can fail, and for this, the above framework more clearly shows the aspects in terms of (I) and (II) that can make these assumptions invalid. This in turn can indicate designing new studies and methods that will put those aspects to the test, to validate or revise the previous results.

Example of explicit assumptions on potential outcomes and principal stratification.

We give meaning to the above loci by the following conditions (see also Fig. 3). The locus for m stands for a marker we measure, which, although can be close to g or z, is itself non-causal for either the trait or the covariate. That is, $Y_i(g, z, m) = Y_i(g, z, m') = Y_i(g, z)$ and $X_i(g, z, m) = X_i(g, z, m') = X_i(g, z)$ (so m is omitted from the argument of the trait and covariate). The locus z (but not g) stands for a genetic determinant of the covariate X, in the sense that $X_i(g', z) = X_i(g, z) = X_i(z)$ and that $X_i()$ is different for at least two genotypes z and z'. For example, alcohol consumption, and smoking, which are often used as covariates in linkage of other traits, have been found themselves to have genetic determinants that predispose to addiction (Wang et al., 2004; Pianezza et al. 1998). We allow that the main trait Y can be affected by the gene g that does not affect the covariate X (thus allowing for a direct genetic effect on Y, in pathways other than through X); and we also allow that Y can be affected by the gene z that affects the covariate X (which allows that Y can depend on X). These conditions *also allow* for other environmental factors to affect the trait or the covariate and to interact with g, z, because we allow the functions Y and X to depend also on the unit i.

General goals in this problem then can be to learn about the locus of the gene g that affects the main trait in pathways other than through X, and the degree of this effect.

For the set of possible g, z, the variability created by the events of parental meiosis and fertilization induces a joint distribution for the set of potential outcomes $\{X_i(z), Y_i(g, z)\}$, denoted by pr$(X_i(z), Y_i(g, z))$. At a particular event of parental meioses and fertilization, a unit i is actually attached to one genotype at each of the three loci, which we denote by G_i, Z_i, M_i respectively. The induced distribution of the functions $\{X_i(z), Y_i(g, z)\}$ and the actual genotypes will be denoted by pr$(\{X_i(z),\ Y_i(g, z)\},\ G_i,\ Z_i,\ M_i)$. The observed covariate status X_i and observed main trait status Y_i of a unit, then, are the result from the coupling of the potential outcomes for that unit and the observed genotypes, that is, $X_i = X_i(Z_i)$ and $Y_i = Y_i(G_i, Z_i)$. For this part, the marker M_i is known but the genotypes G_i, Z_i are not directly observed because the loci are unknown. We consider the following assumptions (conditionally on the parental genotypes).

(A.1). The potential outcomes $\{X_i(z), Y_i(g, z)\}$ are independent of the actual genotypes (G_i, Z_i, M_i). This assumption still allows that the genetic material of a unit at distances close to the loci for g, m, z be correlated to (G_i, Z_i, M_i).

The assumption means that the *potential* of units to develop the covariate and the trait, *as function* of the possible genotypes in the three loci is independent of the actual genotypes at those loci. This assumption is expected to be correct if any genes *other* than g, z that affect Y and X are far from g, z and m. This assumption, though not usually explicit, is implicit and necessary for the interpretation of results in most linkage analyses.

(A.2). To distinguish between the two pathways g, z for Y, we assume that Y is a function of z only through the covariate: if gene z does not affect the covariate of a unit, then it does not affect the main trait, that is, if $X_i(z) = X_i(z')$ then $Y_i(g, z) = Y_i(g, z')$.

(A.3). The potential $Y_i(g, z)$ for getting the trait for fixed genotypes g, z has the same distribution among people with $\{i : X_i(z) = x\}$ as it has among people in the stratum $\{i : X_i(z) = X_i(z') = x, \text{ for all } z'\}$, also known as a principal stratum (Frangakis and Rubin, 2002).

(A.4). The loci for z and g are not close to each other from a perspective of probability of recombination given marker data, in the sense that Z_i and G_i are independent given M_i.

In relation to the above two goals, we focus here on the locus of the gene g that affects the main trait in pathways other than through X. To examine if the gene g is close to the studied marker, we examine testable implications of assuming instead the null hypothesis that this distance is large, namely, that the genotypes G_i and M_i are transmitted independently:

$$H_0 : \text{pr}(G_i, M_i) = \text{pr}(G_i)\text{pr}(M_i). \tag{1}$$

We are also interested in the joint distribution of the trait and covariate status, the genotype of interest, and the marker, which under assumptions A.1-A.4 is obtained as

$$\text{pr}(Y_i, X_i, G_i, M_i) = \text{pr}(Y_i \mid G_i, X_i)\, \text{pr}(X_i \mid M_i)\, \text{pr}(G_i, M_i) \tag{2}$$

(proof omitted, available on request). The above arguments bring up two important issues.

First, in the next section we show that even under these relatively simplified conditions, the regression methods that are routinely used to test the null hypothesis H_0 of expr. (1) of no linkage of the trait are generally inappropriate when, as above, the covariate has a genetic determinant. Under those conditions we also propose methods to address the goals.

Second, note that, although the assumption in (2) is frequently and implicitly assumed in the observed data, it requires generally all of assumptions A.1-A.4 on the underlying mechanisms. For example, even if we assume there is no direct causal effect of the oovariates' gene (z) on the disease except through the covariate, assumption (2) will be false unless assumption A.3 holds. The latter assumption excluded a type of selection bias, and needs to be expressed in terms of strata defined by the joint potential values of the covariate. Such strata are known as principal strata and play a key role in more general studies with multiple causal factors (Frangakis and Rubin, 2002); other applications of principal stratification include studies with noncompliance to treatment, direct and indirect effects, intermediate and surrogate endpoints (e.g., Imbens and Rubin, 1997; Rubin, 2004; Frangakis et al. 2004; Gilbert et al. 2003). Of course, assumptions A.1-A.4 are not necessarily correct. However, failure of the standard methods under these simplifying assumptions implies failure under more general assumptions, so possible violations of A.1-A.4 are not an argument against the comparative advantage of the new methods we will discuss. Moreover, if there is concern on the validity of (2), the above framework allows us to see that the source of that violation can be violation of any one (or combinations) of A.1-A.4, and this can suggest more specific designs and methods to examine them.

3. Problems with standard methods.

We focus on one parental mating type whose genotypes at the marker locus we assume are known and fully polymorphic, so that for every offspring we can measure genetic similarity by the number of alleles identical by descent (IBD) at the marker locus.

Consider first the conditional-logistic regression approach to incorporate covariates into affected sib pairs, proposed by Olson (1999). This approach finds the conditional probability $\hat{f}_{p,k}$, given the data of sib pair p, that the pair shares k alleles IBD, and compares it to the unconditional probability $f_k = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, for $k = 0, 1, 2$, respectively. Here and below the added subscripts $1, 2$ for M_p, X_p, Y_p represent the two member of a sib pair p, and an underline represents both. Since the marker is fully informative, $\hat{f}_{p,k}$ is known precisely for any pair, and is either 0 or 1. Letting X_p^* be a pair-specific function of the two individual covariates $(X_{p,1}, X_{p,2})$ in a pair (e.g., sum, absolute difference, or both values), Olson (1999) considered the likelihood ratio (LR) of observing marker data $\underline{M}_p$ from a pair p given covariate X_p^* and that both members are affected, versus that likelihood given X^* but unconditionally on affection status, as

$$\begin{aligned}\text{LR} &= \frac{\text{pr}(\underline{M}|Y_1 = Y_2 = 1, X^*)}{\text{pr}(\underline{M}|X^*)} \\ &= \frac{\sum_{k=0,1,2}\{\text{pr}(Y_2 = 1|\text{IBD} = k, Y_1 = 1, X^*)/\text{pr}(Y_2 = 1|X^*)\}\text{pr}(\text{IBD} = k|\underline{M}, X^*)}{\sum_{k=0,1,2}\{\text{pr}(Y_2 = 1|\text{IBD} = k, Y_1 = 1, X^*)/\text{pr}(Y_2 = 1|X^*)\}\text{pr}(\text{IBD} = k|X^*)} \qquad (3)\end{aligned}$$

(we have omitted subscript p for brevity). The work by Olson (1999; see also Goddard et. al, 2001; SAGE, 2003) essentially (though implicitly) proposed to parameterize the recurrence risk ratio for one individual who shares IBD=k with an affected sibling and with pair specific covariate X^* as

$$\frac{\text{pr}(Y_2 = 1|\text{IBD} = k, Y_1 = 1, X^*)}{\text{pr}(Y_2 = 1|X^*)} = \exp(\beta_k' + \delta_k' X^*), \qquad (4)$$

In this model, the null hypothesis H_0 of the previous section – that the marker locus is not linked to the gene g that affects Y and not X – was implicitly assumed to be represented by the hypothesis H_0^* that (4) does not depend on the marker's IBD, namely that $H_0^* : \beta_0' = \beta_1' = \beta_2'; \delta_0' = \delta_1' = \delta_2'$. Based on expression (4), they considered the likelihood ratio (3) for a pair to be

$$\mathrm{LR}(\beta, \delta) = \frac{\sum_{k=0,1,2} e^{\beta_k + \delta_k X^*} \hat{f}_k}{\sum_{k=0,1,2} e^{\beta_k + \delta_k X^*} f_k}, \tag{5}$$

where $\beta_0 = 0, \beta_k = \beta_k' - \beta_0'$ and $\delta_0 = 0, \delta_k = \delta_k' - \delta_0'$. This reparametrization, which was also implicit, was considered because not all parameters are identifiable from (5), and implies that:

$$H_0^* : \beta_k = 0 \text{ and } \delta_k = 0, \text{ for all } k. \tag{6}$$

Then, the existing methods (e.g., SAGE, 2003, and references therein) for testing H_0^* estimate the parameters in (4) by the maximum likelihood estimate (MLE) of β_k and $\delta_k (k = 1, 2)$ of the product of LR in (5) over the available ASPs subject to the usual triangle constraints (Holmans, 1993).

In order for the likelihood ratio (5) in Olson (1999) and Goddard (2001) to be derived from the likelihood ratio of the general case (3), we observe that the conditional distribution of alleles shared IBD given covariates X^*, pr$(\text{IBD} = k|X^*)$, is implicitly assumed to equal the unconditional distribution pr$(\text{IBD} = k)$ which is $f_k = (1/4, 1/2, 1/4), k = 0, 1, 2$. However, when the studied marker (among many) comes close to the genetic determinant z of the covariate, this assumption is not necessarily true. The possibility the assumption is violated has been expressed earlier in the literature (e.g., Greenwood and Bull, 1999; Schaid et al., 2003), but its consequences have not been appreciated. For example, Schaid et al. (2003, p. 90) express that such violation has no consequences in analyses if, as is commonly the case, the analyses use the sum $X^* := (X_1 + X_2)$ as the pair-specific covariate. For this reason, the possible consequences of this assumption are ignored. The above approach is now implemented and used as standard by commercial software (SAGE, 2003) and forms a routine tool in research for trait genes.

To our knowledge, however, neither the original nor later research has stated explicit conditions in terms of functional relations (such as those of the previous section) that can make the resulting methods valid for testing the hypotheses H_0 or H_0^*, even for proposals such as using the sum as the pair-specific summary.

4. Proposed design strategy to address the problem.

Closer examination of the role of the above implicit assumption in estimation, when considering specific models such as assumptions A.1-A.4, shows that the above methods are not appropriate. Specifically, under assumptions A.1-A.4, we have the following result.

RESULT. Suppose model (4) holds and the null hypotheses of no linkage H_0 and H_0^* are true. Then:
(a) the MLEs $\hat{\beta}_k$ and $\hat{\delta}_k$ (for $k = 1, 2$) from the product of independent terms (5) will, with increasing number of ASPs, generally converge to non-zero values. This problem also holds if the sum $X^* := (X_1 + X_2)$ is used as the pair specific summary of the covariate.
(b) the null hypothesis H_0 has implications testable from family data.

This result means that loci identified by the standard approach as linked to the main disease trait conditionally on the covariate, can, in reality, be linked only to the covariate and can have absolutely no linkage to the trait. We prove the above result by giving an example where the trait is not even related to the covariate.

Proof by example. Consider the following conditions. The gene z for a binary covariate X has four alleles, $\zeta_1, ..., \zeta_4$, and the gene g for the separate pathway to trait Y has two alleles, (T, t); for the covariate, pr$(X_i(z) = 1) = 10\%$ if none of the two alleles of z is ζ_1, but

$\text{pr}(X_i(z) = 1) = 30\%$ if any of the alleles in z is ζ_1; the trait is not affected by z at all; $\text{pr}(Y_i(g) = 1) = (20\%, 10\%, 5\%)$, respectively, for $g = TT, Tt$ and tt; and $Y_i(g)$ and $X_i(z)$ are independent. To make the marker close to the genetic determinant of X, suppose that the marker *is* at z. Siblings are to be studied at the marker from families, say, with a type of parents being $\zeta_1 T/\zeta_2 t \times \zeta_3 T/\zeta_4 t$. We record the families with two affected children, and focus on the pairs where the sum variable $X^* = X_1 + X_2$ is either =0 or 2. Finally, assume the null hypothesis H_0 is true for all families.

With these conditions, the model (4) is correct for all families (as was intended to be), with all $\beta'_k = \delta'_k = 0$ and so all $\beta_k = \delta_k = 0$, because in (2) the trait now is independent of both the marker data and the covariate. However, the MLEs of the parameters using (5) can be shown to converge to plim$\hat{\beta}_1 = 0.33,$ plim$\hat{\beta}_2 = 0.49,$ plim$\hat{\delta}_1 = -0.38,$ plim$\hat{\delta}_2 = -0.59$, which happens because stratification on X increases sharing at the marker. From the interpretation of methods as currently used, these results would suggest that at the covariate level $X^* = 0$, i.e., when both members of the pair have covariate 0, the marker is close to a genetic determinant of the trait, since the recurrence risk ratio (4) would be estimated as 1.39 and 1.63 for IBD=1 and 2 vs. 0. These results and interpretations, as we see now, would be misleading since, in reality, the marker is completely unlinked to the trait's determinant.

This problem is not limited to the above conditional logistic model approach. For example, we can show that the mixture model approach proposed by Devlin et. al (2002) suffers from the same problem (proof available on request).

Better methods to address this problem need to explicitly allow for a model where the covariate(s) can have genetic determinants. When such a model is posited, we can then investigate testable implications of the null hypothesis H_0.

An example of such an explicit model is the one under assumptions A.1-A.4. Under that model, we can show that if the null hypothesis H_0 is true, then given covariate values $\underline{X}$ of a *general* sib-pair (i.e., not necessarily affected), the distribution of alleles shared IBD: (i) is not necessarily the unconditional null distribution $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$; but (ii) is, nevertheless, independent of the pair's disease status $\underline{Y}$:

$$\begin{aligned} \text{pr}(\underline{Y}|\underline{M}, \underline{X}) &= \sum_{\underline{g}} \text{pr}(\underline{Y}|\underline{G} = \underline{g}, \underline{M}, \underline{X})\text{pr}(\underline{G} = \underline{g}|\underline{M}, \underline{X}) \\ &= \sum_{\underline{g}} \text{pr}(\underline{Y}|\underline{G} = \underline{g}, \underline{X})\text{pr}(\underline{G} = \underline{g}|\underline{X}) = \text{pr}(\underline{Y}|\underline{X}). \qquad (7) \end{aligned}$$

Relation (7), therefore, provides a testable implication of the hypothesis H_0, proving part (b) of the above Result, and it also indicates that more than affected-sib-pairs are needed for such methods. Specifically, for a stratum of pair-wise covariates $\underline{X}$, consider the following table, where entry $p_{k,\underline{Y}}$ represents the probability that a pair with disease status $\underline{Y}$ shares k alleles IBD.

$\underline{X}$	$\underline{Y}$	IBD=0	IBD=1	IBD=2
(X_1, X_2)	$(0, 0)$	$p_{0,00}$	$p_{1,00}$	$p_{2,00}$
	$(0, 1)$	$p_{0,01}$	$p_{1,01}$	$p_{2,01}$
	$(1, 0)$	$p_{0,10}$	$p_{1,10}$	$p_{2,10}$
	$(1, 1)$	$p_{0,11}$	$p_{1,11}$	$p_{2,11}$

In the above table, then, expression (7) implies that if H_0 is true, the row probabilities must all be the same. If, on the other hand, the marker is linked to g, both ASPs and unaffected sib pairs (USPs) are expected to have increased sharing since both are similar in their trait status. So, the distributions of IBD should differ to the largest degree for ASPs (and USPs) versus discordant sib pairs (DPSs), that is, between the first (and fourth) versus the second and third rows of the above table. Therefore, substantial discrepancy in distributions of IBD alleles between ASPs and DSPs (and not between ASPs and $(\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$) with the same covariates $\underline{X}$ is evidence for linkage between the marker and the trait.

This approach also opens the way for considering a number of practical issues. First, because existing studies usually do not collect good quality DNA data for unaffected sib-pairs, they have insufficient information to apply the above methods. Therefore, the above arguments and methods emphasize the importance of collecting such data. Second, in practice each family will be able to provide data on some but not all of the cells in the above table. Therefore, the above approach needs to be modified to be able to use collectively data with continuous covariates and few individuals in each family, for example, with generalized linear models with conditional inference.

In addition, it will be important to modify the above approach to enable it, more generally than testing the null H_0, to estimate physical location of the genes that affect the disease in ways other than through the covariate. This can be achieved by using, as for other questions, multipoint marker information. It will also be important to make fuller use of data than the pair-wise structure usually used.

5. Conclusion.

The proposed causal framework explicitly shows the different types of assumptions, as with assumptions A.1-A.4, that are needed for inference when studying genes that can affect a disease through covariates, together with genes that can affect a disease in ways other than through the covariates. The framework shows that even under relatively simple assumptions, standard methods do not address the goals, yet information about these goals still exist. Therefore, this framework, more generally, allows the development of designs and analyses with validity under explicitly postulated assumptions, and also the testing of implications of such assumptions, thus helping to confirm them, or, alternatively, to point to alternative mechanisms for understanding and ultimately dealing with disease.

APPENDIX

To show that under assumptions A.1-A.4 the MLE of the conditional logistic model is generally inconsistent for the null values of the parameters, we need only show this for a model that saturates a discrete covariate, say a binary X^*. To represent such a setting, focus on independent sib pairs p whose members have either both their covariate value 0, in which case we now label $X_p^* = 0$, or both their covariate value 1, in which case we now label $X_p^* = 1$.

The maximum likelihood estimators maximize the average, say $Q_n(\beta, \delta)$, of the log of the product of terms in (5), that is,

$$Q_n(\beta, \delta) \quad = \quad \frac{1}{n} \sum_{p=1}^{n} \log \frac{\sum_{k=0,1,2} e^{\beta_k + \delta_k X_p^*} \hat{f}_{p,k}}{\sum_{k=0,1,2} e^{\beta_k + \delta_k X_p^*} f_k}$$

where $\hat{f}_{p,k} = I(\text{IBD}_p = k)$. By the weak law of large numbers, and since the numerator involves in fact only one term, we have that

$$\begin{aligned} Q_n(\beta, \delta) \quad &\to_p \quad E\left\{ \log \frac{\sum_{k=0,1,2} e^{\beta_k + \delta_k X_p^*} \hat{f}_{p,k}}{\sum_{k=0,1,2} e^{\beta_k + \delta_k X_p^*} f_k} \right\} \\ &= \quad \sum_{k,x^*} (\beta_k + \delta_k x^*) \text{pr}(\text{IBD} = k, X^* = x^*) - \sum_{x^*} \log\{\sum_k e^{\beta_k + \delta_k x^*} f_k\} \text{pr}(X^* = x^*). \end{aligned}$$

Let $q(\beta, \delta)$ denote the last expression. Under general conditions, the probability limits, say $\tilde{\beta}$ and $\tilde{\delta}$, of the MLE of β and δ are the maximizers of $q(\beta, \delta)$, that is, they solve

$$\frac{\partial q(\beta, \delta)}{\partial \beta_k} = 0, \qquad \frac{\partial q(\beta, \delta)}{\partial \delta_k} = 0, \quad k = 1, 2.$$

Let g_{kx^*} denote the conditional probabilities of IBD sharing, pr(IBD $= k | X^* = x^*$). It is easy then to show that

$$\tilde{\beta}_k = \log(g_{k0}f_0/g_{k0}f_i); \quad \tilde{\delta}_k = \log(g_{k1}f_0/g_{k1}f_k) - \log(g_{k0}f_0/g_{00}f_k)$$

for $k = 1, 2$. Now, if the marker where IBD is measured is close to genes related to X^*, then pr(IBD $= k|X^* = x^*$) will differ from the unconditional IBD frequencies f_k, so g_{kx^*}/f_k will not equal 1. Therefore, the limit parameter values in the last expression will differ from 0.

Acknowledgement

We thank Dr. Joan Bailey Wilson for useful comments, and the National Eye Institute and the National Human Genome Research Institute for partial support.

REFERENCES

[1] Devlin B, Jones BL, Bacanu SA, Roeder K. (2002) Mixture models for linkage analysis of affected sibling pairs and covariates. *Genet Epidemiol* **22**, 52–65.

[2] Frangakis, CE, and Rubin, DB (2002). Principal stratification in causal inference. *Biometrics*, **58**, 21–29.

[3] Frangakis, CE, Brookmeyer, RS, Varadhan, R, Mahboobeh, S, Vlahov, D, and Strathdee, SA. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a Needle Exchange Program. *Journal of the American Statistical Association* **99**, 239–249.

[4] Gilbert, P. B., Bosch, R. J., and M. G. Hudgens (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in AIDS vaccine trials. *Biometrics*, **59**, 531–541.

[5] Goddard KA, Witte JS, Suarez BK, Catalona WJ, Olson JM. (2001). Model-free linkage analysis with covariates confirms linkage of prostate cancer to chromosomes 1 and 4. *Am J Hum Genet* **68**, 1197-1206.

[6] Greenwood CM, Bull SB. (1997). Incorporation of covariates into genome scanning using sib-pair analysis in bipolar affective disorder. *Genet Epidemiol* **14**, 635–640.

[7] Greenwood CM, Bull SB. (1999). Analysis of affected sib pairs, with covariates–with and without constraints. *Am J Hum Genet* **64**, 871–885.

[8] Holmans P. (1993). Asymptotic properties of affected-sib-pair linkage analysis. *Am J Hum Genet.* **52** 362-374.

[9] Imbens, G. W. and Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* **25**, 305–327.

[10] Mirea L, Briollais L, Bull S. (2004). Tests for covariate-associated heterogeneity in IBD allele sharing of affected relatives. *Genet Epidemiol* **26**, 44–60.

[11] Olson JM. (1999). A general conditional-logistic model for a affected-relative-pair linkage studies. *Am J Hum Genet* **65**, 1760–1769.

[12] Olson JM, Song Y, Dudek DM et al. (2002). A genome screen of systemic lupus erythematosus using affected-relative-pair linkage analysis with covariates demonstrates genetic heterogeneity. *Genes Immun* **3** Suppl 1: S5-S12.

[13] Penrose LS. (1935). The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Ann Eugen* **6**, 133–138.

[14] Pianezza ML, Sellers EM, Tyndale RF. (1998). Nicotine metabolism defect reduces smoking. *Nature* **393**, 750.

[15] Risch (1990a). Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected relative pairs. *Am J Hum Genet* **46**, 242–253.

[16] Risch (1990b). Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet. **46**, 229-241.

[17] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66**, 688–701.

[18] Rubin, D. B. (1978). Bayesian inference for causal effects. *Annals of Statistics* **6**, 34–58.

[19] Rubin (2004). Direct and Indirect Causal Effects Via Potential Outcomes. *Scandinavian Journal of Statistics* **31**, 161-170 (with discussion).

[20] S.A.G.E. Statistical Analysis for Genetic Epidemiology. (2003). Computer program package available from Statistical Solutions Ltd, Cork, Ireland.

[21] Schaid DJ, Olson JM, Gauderman WJ, Elston RC. 2003. Regression models for linkage: issues of traits, covariates, heterogeneity, and interaction. *Hum Hered* **55**, 86–96.

[22] Wang JC et al. (2004). Evidence of common and specific genetic effects: association of the muscarinic acetylcholine receptor M2 (CHRM2) gene with alcohol dependence and major depressive syndrome. *Hum Mol Genet.* **13**, 1903-1911.

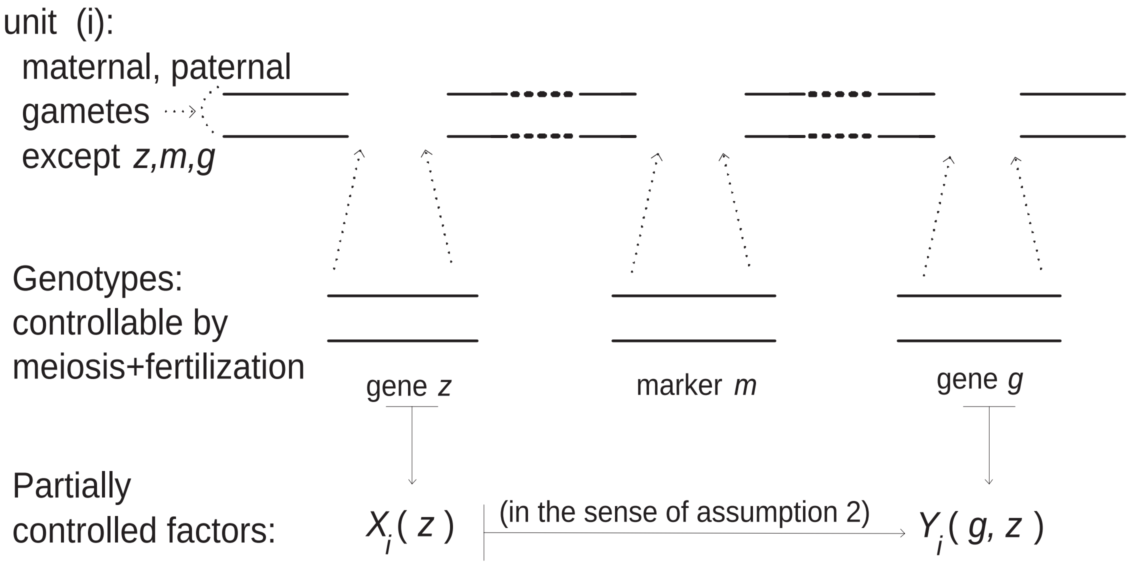

Figure 1. Linkage of a trait to genetic locus and covariate with genetic determinant.