Genetic Epidemiology 22:52–65 (2002)

Mixture Models for Linkage Analysis of Affected Sibling Pairs and Covariates

B. Devlin,[1]* Bobby L. Jones,[2] Silviu-Alin Bacanu,[1] and Kathryn Roeder[2]

[1]*Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania*
[2]*Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania*

To determine the genetic etiology of complex diseases, a common study design is to recruit affected sib/relative pairs (ASP/ARP) and evaluate their genome-wide distribution of identical by descent (IBD) sharing using a set of highly polymorphic markers. Other attributes or environmental exposures of the ASP/ARP, which are thought to affect liability to disease, are sometimes collected. Conceivably, these covariates could refine the linkage analysis. Most published methods for ASP/ARP linkage with covariates can be conceptualized as logistic models in which IBD status of the ASP is predicted by pair-specific covariates. We develop a different approach to the problem of ASP analysis in the presence of covariates, one that extends naturally to ARP under certain conditions. For ASP linkage analysis, we formulate a mixture model in which a disease mutation is segregating in only a fraction α of the sibships, with $1 - \alpha$ sibships being unlinked. Covariate information is used to predict membership within groups; in this report, the two groups correspond to the linked and unlinked sibships. For an ASP with covariate(s) $Z = z$ and multilocus genotype $X = x$, the mixture model is $\alpha(z)g(x; \lambda) + [1 - \alpha(z)]g_0(x)$, in which $g_0(x)$ follows the distribution of genotypes under the null IBD distribution and $g(x; \lambda)$ allows for increased IBD sharing. Two mixture models are developed. The pre-clustering model uses covariate information to form probabilistic clusters and then tests for excess IBD sharing independent of the covariates. The Cov-IBD model determines probabilistic group membership by joint consideration of covariate and IBD values. Simulations show that incorporating covariates into linkage analysis can enhance power substantially. A feature of our conceptualization

Contract grant sponsor: National Institutes of Health; contract grants: MH57881 and MH56193; Contract grant sponsor: NSF; contract grant: DMS 9803433.

B.L. Jones is currently at Department of Human Genetics, University of Pittsburgh, Pittsburgh, Pennsylvania.

*Correspondence to: Bernie Devlin, Department of Psychiatry, University of Pittsburgh School of Medicine, 3811 O'Hara Street, Pittsburgh, PA 15213. E-mail: devlinbj@msx.upmc.edu

Received for publication 13 March 2001; revision accepted 16 April 2001

of ASP linkage analysis, with covariates, is that it is apparent how data analysis might evaluate covariates prior to the linkage analysis, thus avoiding the loss of power described by Leal and Ott [2000] when data are stratified. Genet. Epidemiol. 22:52–65, 2002.

Key words: clustering algorithms; mixing distribution; score statistics; likelihood ratio; asymptotic distributions

INTRODUCTION

Risch [1990a,b,c], in a seminal series of papers, laid out the theoretical basis and empirical advantages of affected sib-pair (ASP) and affected relative pair (ARP) methods for linkage analysis of complex diseases. Since then, many ASP/ARP studies have been initiated to determine the genetic underpinnings of complex diseases. Unfortunately, many complex diseases have turned out to be complex indeed, and most of these studies produced ambiguous results. ASP/ARP methods, although inherently robust to different disease models [Risch, 1990b], are not necessarily powerful approaches for disease gene mapping. Consequently auxiliary data to improve power should be sought whenever possible.

An excellent example of critical auxiliary data is age at onset for both breast cancer and Alzheimer's disease. Both of these diseases have early-onset forms that are strongly genetic, and late-onset forms that are determined by a combination of environmental and genetic factors. If one were to analyze a cohort of ASPs who have breast cancer or Alzheimer's disease for linkage, then age at onset would be a critical covariate. In fact, if age at onset had not been taken into account, linkage to *BRCA1* would have been missed because the overall LOD score was negative [Hall et al., 1990].

Type II diabetes and prostate cancer offer examples in which covariates may be critical. For type II diabetes, Ghosh et al. [2000] utilized fasting insulin, fasting glucose, and body mass index when performing an ordered-subsets analysis [Hauser et al., 1998] of genome scan data. For prostate cancer, Goddard et al. [2001] utilized Gleason score, age at onset, male-to-male transmission, and number of first-degree relatives to enhance their power to detect linkage (see also Schaid et al. [2001]).

In addition to the ordered-subsets analysis [Hauser et al., 1998], there are numerous other methods available to analyze ASP linkage data with covariates. The most recent work traces to Rice [Dorr et al., 1997; Rice et al., 1999], Greenwood and Bull [1997, 1999a], Olson [1999], and Goddard et al. [2001]. Their models share the underlying feature that IBD status of the ASP is regressed against the covariates; in the Discussion section, we contrast the essential features of our models with their recent work.

The framework for our models, which are developed for ASP but extend to ARP under certain conditions, can be understood by appealing to pedigree analysis under a genetic heterogeneity model [Smith, 1963; Ott, 1983; Matise and Weeks, 1993]. Under heterogeneity, it is assumed a certain portion α of the pedigrees demonstrate linkage between markers in the vicinity of a disease gene. The remaining fraction of families, $1 - \alpha$, have the disease for some other reason. For a polymorphic marker in the region of interest, the likelihood of a set of pedigrees is maximized with respect to the recombination fraction θ and the heterogeneity parameter α.

For complex disease, we assume ASP in the sample differ with respect to sources of liability. Some ASP share a particular source of liability in common—polymorphisms in the same gene—whereas other ASP do not share that source of liability. By analogy to the linkage heterogeneity model, those ASP who share a common source of liability are "linked," whereas the remainder are "unlinked." Covariates can convey information regarding which ASP have common sources of liability and therefore are more likely to be in the linked group. This information can be incorporated into a test for linkage using the linkage heterogeneity framework. All ASP are used in the analysis, but their contribution to the likelihood of linkage is weighted by their probability of being in the linked group. Based on this principle, Schaid et al. [2001] develop a parametric mixture model; we develop a similar approach that utilized a non-parametric mixture model.

By analogy to the analysis of breast cancer ASP, age at onset for each ASP largely would determine group membership to either the linked or unlinked group; conditional on group membership, the IBD distribution follows the null distribution (unlinked group) or an alternative distribution exhibiting enriched IBD sharing (linked group). Each ASP is assigned a probability of group membership via the mixing parameter $\alpha(Z)$. When examining a locus that is unlinked, the parameter of the linked distribution can default to the unlinked status, even if the covariates separate the families into two clusters.

To motivate our model, we continue with the example of age at onset for breast cancer. As described more fully later, we have generated a portion α of ASP that exhibit age at onset like that of families carrying a *BRCA1*-like mutation; for the remaining $(1 - \alpha)$ ASP, the age at onset is similar to that for families not carrying early-onset mutations. From these data, we drew 200 ASP at random from our simulated data, with each ASP meeting the condition that the age at onset of the proband be ≤65 years. Plotting the ASP's ages at onset against each other, with the smaller value on the abscissa (Fig. 1a), we can see some tendency for ASP to cluster at earlier ages at onset. The clustering is made more obvious by the symbols, which denote ASP sharing of a *BRCA1*-like mutation by a +. Simple cluster analysis yields two intuitive clusters (labeled * and o in Fig. 1b). For the total of 200 ASP (Fig. 1a), 59 or 29.5% share a mutation at the *BRCA1*-like gene. For the cluster labeled by * in Fig. 1b, which is the cluster of interest based on the biology of breast cancer [Claus et al., 1990], the fraction of ASP sharing the mutation is 47.3% of 110 ASP. This cluster is more homogeneous and more likely to yield a noteworthy linkage signal. In fact, it contains all but nine of the ASP that share the *BRCA1*-like mutation. Clearly age at onset, the covariate, provides a powerful signal regarding membership in the linked and unlinked group, and this is the kind of information our mixture models exploit.

In this paper, we introduce two mixture models for ASP linkage analysis with covariates. One model completely ignores IBD status of the ASP, using only the covariate information for clustering; the other model uses both sources of information. Both models use multipoint IBD information for linkage by using the output from convenient linkage packages such as GENEHUNTER [Kruglyak et al., 1996] and ASPEX [Jorgenson et al., 1999]. Given this IBD information, implementing the models is straightforward and yields quite powerful inference when the covariates carry information about membership into linked and unlinked groups.

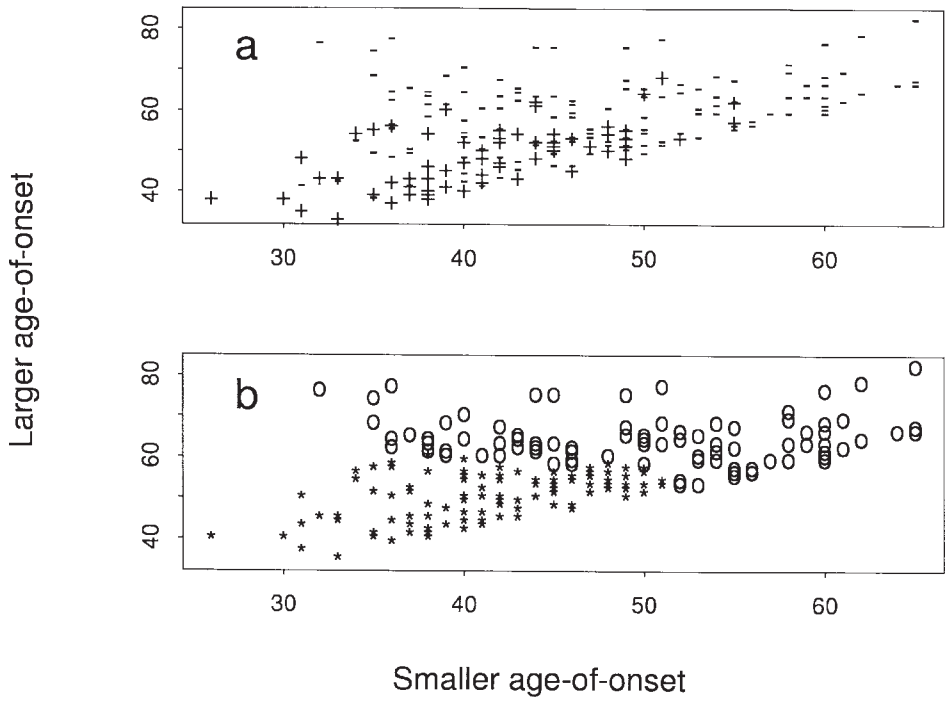

Fig. 1. Simulated age at onset data (the covariate) for affected sib pairs (ASP) who present with a breast cancer–like disease. **a**: + represents ASP due to a *BRCA1*-like mutation; – represents sib pairs affected due to other causes. **b**: Clustering of ASP into two groups by age at onset, in which * equals the early age at onset group.

DEFINITION OF NOTATION AND LIKELIHOOD

In this article, we focus on ASP, which are assumed to be independent. In a future article, we plan to extend the model to more complex family structures. Suppose n families are collected, each with a single ASP. For each pair, we observe multilocus marker data X_i, typically from a genome scan, and a set of covariates Z_i, so that the completely observable data are (X_i, Z_i), $i = 1, ..., n$. With parents genotyped for a set of fully informative markers spanning position d, IBD sharing Y_i at position d is also observable; however, the Y_i are treated as unobservable because the requirement of fully informative, dense markers is generally not met. Write $f_0(y)$ for the trinomial model for IBD sharing under the null hypothesis of no linkage, which has probabilities of (1/4, 1/2, 1/4) mapping onto IBD status (0, 1, 2), respectively. Consistent with the Risch [1990b] parameterization, we write the trinomial under the alternative model of linkage as $f(y; \lambda)$ with probabilities $[1/(4\lambda), 1/2, 1/2 - 1/(4\lambda)]$ mapping onto (0, 1, 2) IBD, respectively [Holmans, 1993; Kong and Cox, 1997]. Here λ corresponds to Risch's [1990a] λ_s, the recurrence risk ratio for the sibling of an affected individual, defined additively. When placed in the mixture model framework we develop herein, it has the advantage of meeting the usual triangle constraints for ASP allele sharing [Holmans, 1993].

Assume a polymorphism at position d has an impact on liability for disease and on covariate values and that a fraction of the sample has the disease due to the effect of the polymorphism (in part or in total). As described in the introduction, we refer to the ASP as linked when $C = 1$ and unlinked when $C = 0$. Clearly C is unobservable.

For a particular ASP, define $\alpha(z)$ as the probability that C =1 based on the observed covariates Z. The likelihood of observing $Y = y$ for a sib pair with covariates $Z = z$ is $h[y; \lambda, \alpha(z)] = \alpha(z) f(y; \lambda) + [1 - \alpha(z)] f_0(y)$.

The likelihood for the observable quantity is $p[x; \lambda, \alpha(z)] = \Sigma_y Pr(X = x \mid Y = y) h[y; \lambda, \alpha(z)]$. The weighting function $Pr(X = x \mid Y = y)$ can be obtained using various programs, such as GENEHUNTER [Kruglyak et al., 1996] and ASPEX [Jorgenson et al., 1999], which compute the reverse conditional, $p(y \mid x)$, assuming $\lambda = 1$ [see Fig. 1 in Kruglyak et al., 1996]. Thus, the desired quantity is obtained via Bayes theorem $p(X = x \mid Y = y) = p(y \mid x) g_0(x)/f_0(y)$, where $g_0(x)$ denotes the marginal probability of observing the multilocus genotype configuration $X = x$, assuming $\lambda = 1$. Now the full likelihood can also be expressed as a mixture model, $p[x; \lambda, \alpha(z)] = \alpha(z) g(x; \lambda) + [1 - \alpha(z)] g_0(x)$, in which $g(x; \lambda) = g_0(x) \Sigma_y p(y \mid x) f(y; \lambda)/f_0(y)$.

We consider two approaches to estimation and testing. In one approach, α is estimated using the covariates only, and clusters are determined without regard for IBD information. We call this the pre-clustering model because weights for membership into the linked/unlinked clusters are determined before any linkage analysis. Alternatively, the likelihood could be jointly maximized with respect to weights for group membership and IBD sharing. We call this the cov-IBD model.

TESTING FOR LINKAGE

Before discussing specific models and tests, it is useful to discuss the distribution of test statistics for mixture models and how it applies to our linkage setting. Testing for linkage is equivalent to testing $\lambda = 1$. Define the log-likelihood as $l[\lambda, \alpha(z)] = \Sigma_i \log p[x_i; \lambda, \alpha(z_i)]$. Under the null hypothesis the log-likelihood reduces to $l(1) = \Sigma_i \log g_0(x_i)$. Notice that the log-likelihood ratio,

$$LR[\alpha,\lambda] = \sum_i \log\left\{\alpha(z_i)\left[\sum_y p(y \mid x_i)\frac{f(y;\lambda)}{f_0(y)}\right] + [1-\alpha(z_i)]\right\}, \quad (1)$$

does not depend on $g_0(x)$.

Let $\hat{\lambda}$ be the maximum likelihood estimate for λ, which is constrained to be greater than or equal to 1. If $\alpha(z)$ were known, then the likelihood-ratio test is approximately distributed as a one-sided χ^2_1 test, i.e., a 1/2:1/2 mixture of a χ^2_0, which is a point mass at 0, and a χ^2_1. When $\alpha(z)$ is unknown, the likelihood ratio test does not follow the simple limiting distribution given above; the actual null distribution involves the supremum of a Gaussian process [Ghosh and Sen, 1985; Chakravarti et al., 1987; Chernoff and Lander, 1995]. In practice, the test has low power to detect deviations from the null hypothesis.

A similar problem arises in the context of the linkage heterogeneity model. Both Liang and Rathouz [1999] and Lemdoni and Pons [1995] offer solutions to obtain

simple limiting distributions. Liang and Rathouz [1999] achieve this end by simply pre-specifying the quantity $\alpha = \tilde{\alpha}$, based on considerations external to the linkage distribution. Lemdoni and Pons [1995] note that the likelihood ratio test is well behaved provided α does not approach zero. To avoid this problem, they bound α from below by an arbitrary constant. Even with the constraint, notice that the data can still support the null hypothesis when $\hat{\lambda} \approx 1$.

PRE-CLUSTERING MODEL

In this section, assume that weights for group membership are determined prior to any linkage analysis. To obtain a simple test statistic, Liang and Rathouz [1999] propose a novel score test in which they replace α by an arbitrary pre-specified quantity, $0 < \tilde{\alpha} \leq 1$. In a similar spirit, we propose a related likelihood ratio test with stochastic weights $\tilde{\alpha}(z_i)$ that depend on the covariates z_i. The likelihood ratio test, $\tilde{\Lambda} = 2LR[\tilde{\alpha}; \tilde{\lambda}]$, is evaluated at $\tilde{\lambda}$, the quantity that maximizes equation (1) with $\tilde{\alpha}(z_i)$ replacing $\alpha(z_i)$. Under the null hypothesis, $\tilde{\Lambda}$ also has the one-sided χ^2_1 limiting distribution given above (see Appendix).

Precomputed Clusters

We assume that one or more covariates convey information about membership in the linked and unlinked groups. Clustering ASP into groups could be performed automatically by statistical algorithms, but we believe clustering should be based on biological insights and exploratory data analysis combined with formal statistical clustering methods. The biology of the disease determines the interesting target cluster for linkage analysis. Regardless of the clustering method employed, the investigator chooses which cluster is likely to identify the linked group. Then $\tilde{\alpha}(z_i)$ is an estimator of the probability the *i*th individual is a member of the cluster of interest.

Statistical packages, such as Splus, provide programs for clustering and probabilistic assignment of ASP to groups. For instance, in Splus, one could use *mclust*, which has the functions *mhtree* for clustering and *me* for probability of group membership [Banfield and Raftery, 1993; Fraley and Raftery, 1998]; alternatively, Fuzzy Analysis, or *FANNY*, provides probability of group membership [Kaufman and Rousseeuw, 1990]. Although multiple groups may exist in the data set, the ultimate goal of the analysis is to assign probabilities or weights of membership into the linked and unlinked groups.

COV-IBD MODEL

With a model-based clustering approach, one could jointly maximize the likelihood with respect to the mixing distribution and the IBD distribution. For the linkage heterogeneity model, Lemdoni and Pons [1995] developed a modified likelihood-ratio test in which $l(\lambda, \alpha)$ is maximized over (λ, α) with the constraint that α is bounded from below by an arbitrary small positive constant. Given this slight modification to the usual likelihood-ratio test, these authors demonstrate that their test possesses the one-sided χ^2_1 limiting distribution given above.

When extended to the setting of ASP with covariates, the limiting distribution becomes somewhat more complicated. As in the linkage heterogeneity model of Schaid

et al. [2001], let $\alpha(z; \beta)$ be the logistic probability that $C = 1$ based on the (possibly vector valued) covariates, Z, and parameters β, i.e., $\alpha(z_i; \beta) = \exp(\beta^T z)/[1 + \exp(\beta^T z)]$.

The full log-likelihood is a function of two sets of parameters: $l(\lambda, \beta) = \Sigma_i \log\{\alpha(z_i; \beta)g(x_i; \lambda) + [1 - \alpha(z_i; \beta)]g_0(x_i)\}$. Define $(\hat{\lambda}, \hat{\beta})$ as the parameters that jointly maximize the likelihood with the constraint $\hat{\lambda} \geq 1$. The idea behind this approach is that the covariates serve to define a partition of the families into two sets: one that fits the null distribution $g_0(\cdot)$ and one that permits an excess of IBD sharing as defined by $g(\cdot; \lambda)$. From this viewpoint, it is clear that the pre-clustering and cov-IBD approaches are somewhat different, but aim toward the same purpose: the former clusters the covariates, and the latter clusters the unobservable responses using the covariates to predict this partition.

Define $\hat{\Lambda} = 2\{l(\hat{\lambda}, \hat{\beta}) - l(1)\}$ as the likelihood-ratio test in which α is bounded from below. Based on our simulation study, it appears that the distribution of $\hat{\Lambda}$ can be well approximated in the tails by a mixture of χ^2 distribution, provided $\alpha(z; \hat{\beta})$ is constrained to be greater than some arbitrary small constant. For one covariate, a reasonable approximation is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_2^2$ for tail probabilities 0.05–0.0001. Likewise, for two covariates, a reasonable approximation is $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_3^2$ for tail probabilities 0.05–0.0001. More accurate estimates of the *P*-values require producing simulated data under the null (SIMULATE, see Acknowledgments section for URL), while leaving the covariates unchanged.

Although both the pre-clustering model and the cov-IBD models rely on estimated weights, the pre-clustering model determines these weights without the benefit of the IBD status. Consequently, a bonus from the cov-IBD model is that it provides estimates of the effect of covariates that are associated with the linked group. Thus, it can be used as a discovery tool for risk factors that are genetically based as well as a discovery tool for the genetic factor itself. See Roeder et al. [1999] and Jones et al. [2000] for an application of models of this general form in a different context. On the other hand, the pre-clustering method utilizes the covariate information without the cost of any extra degrees of freedom (DF).

Because the cov-IBD model requires the joint maximization of several parameters, this method will potentially be subject to numerical instability unless the user exercises care. See Finch et al. [1989] for a discussion of the inherent difficulties in maximizing the likelihood for mixture models.

SIMULATIONS

This section illustrates the performance of our method and how it changes with covariate information. To create a heuristic example, we generate data that mimic breast cancer families in some of its features. In particular, we generate a portion α of ASP that exhibit age at onset like that of families carrying a *BRCA1* mutation; for the remaining $(1 - \alpha)$ ASP, the age at onset is similar to that of families not carrying early-onset mutations.

To generate ASP due to a *BRCA1*-like mutation or other causes, we first create parents who carry the mutation with probability equal to 0.006. We then generate a female sib pair with appropriate Mendelian transmission of the *BRCA1*-like mutation, if present, and a tightly linked, completely informative marker. IBD status of the sib pair at the marker is therefore completely known. The proband is chosen

randomly from the sib pair. The youngest sibling is assigned an age between 20 and 80 years, using a uniform probability distribution. The other sibling is 2 to 6 years older; again, the age difference is randomly generated using a uniform distribution. The pairs' affection status and the age at onset are determined using cumulative life probabilities of affection provided by Iverson et al. [2000] in their Fig. 1. Sibling pairs are selected if they are ASP and the proband's age of onset is less than U years, where U varied from 35 to 65 by intervals of 10. For each ASP, the IBD status, the age at onset, and the *BRCA1* carrier status are known. Because we select on the age of the proband, we use the age of the non-proband as the covariate.

To evaluate the effect of utilizing covariates with the pre-clustering method, we compare this test to a test using all the observations (i.e., the pre-clustering test using a weight of one for all ASP). For each of the four scenarios, we simulate 5,000 data sets of sample size n = 100 and 200 ASP. We use *mclust* to cluster the age at onset data and to estimate the probability that each observation is in the cluster defined by younger age at onset. Clearly the covariates substantially enhance the power of the test to detect linkage (Table I). From the median estimated λ for each analytical

TABLE I. Power Comparisons for Likelihood Test Without Preclustering (I), Likelihood-Ratio Test with Pre-clustering (II) and Cov-IBD Test (III)

Age at onset[a]	Method	Median *P*-value	Median λ	Power level = 0.01	Power level = 0.001
		n = 100			
35	I	0.000	2.17	0.952	0.798
	II	0.000	4.09	0.966	0.849
	III	0.000	6.44	0.971	0.872
45	I	0.002	1.69	0.719	0.402
	II	0.000	2.88	0.831	0.563
	III	0.000	5.52	0.830	0.553
55	I	0.016	1.44	0.428	0.165
	II	0.005	2.17	0.610	0.302
	III	0.005	5.44	0.594	0.283
65	I	0.051	1.30	0.238	0.067
	II	0.017	1.83	0.415	0.157
	III	0.015	4.62	0.435	0.160
		n = 200			
35	I	0.000	2.14	0.999	0.994
	II	0.000	4.22	1.000	0.998
	III	0.000	5.96	1.000	0.998
45	I	0.000	1.67	0.964	0.842
	II	0.000	2.82	0.989	0.938
	III	0.000	5.41	0.992	0.945
55	I	0.001	1.44	0.771	0.478
	II	0.000	2.16	0.912	0.711
	III	0.000	5.33	0.925	0.744
65	I	0.010	1.30	0.497	0.207
	II	0.001	1.84	0.749	0.459
	III	0.001	4.45	0.735	0.495

[a]Maximum age at onset of proband.

method, it is also clear that the increased power obtains by increasing the weight attributed to those sib pairs more likely to carry the *BRCA1*-like mutation.

We also test the cov-IBD method on these data sets. For numerical stability, the age at onset covariate is standardized by subtracting the mean age for the sample and dividing by the standard deviation. To compute $\hat{\Lambda}$, the intercept and slope for the logistic function of age, which determine group membership, are bounded by (–6 to 6) and (–4 to 0), respectively, to ensure that the probability of group membership is for the linked group. These intervals also bound $\alpha(z_i; \beta)$ away from zero. Our results (Table I) show the cov-IBD model can be slightly more powerful than the pre-clustering method, apparently by extracting additional information about group membership from the IBD status of the ASP.

To evaluate the type-I error rate of the tests, we simulate 10,000 data sets of sample size $n = 100$ under the null hypothesis of no linked families. The pre-clustering method clearly attains its asymptotic distribution and the cov-IBD method is slightly conservative (Table II).

DISCUSSION

We describe methods to account for covariate information in ASP linkage analysis. Paraphrasing from the pioneering work of Dawson et al. [1990] on the same subject, the principle underlying our proposed methods is that if ASP are drawn from the same population with respect to disease etiology and if the covariate(s) carry information regarding etiology, then the covariate values of the pairs should be similar. In this instance, the covariate information can be used to discover group membership. We assume that there are two groups to be discovered: ASP who have the disease in part or in total due to a polymorphism linked to a tested marker (linked group) and ASP who have the disease for some other reason (unlinked group).

When evaluating traits as covariates, researchers may intuitively look for a variable that demonstrates a large correlation between ASP. Although this correlation can be important, our methods focus on a different feature of the data: Does the covariate provide evidence for heterogeneity within the collection of ASP? In fact, our methods will only prove useful when the ASP are drawn from distinct populations, due to different etiologies, that are partially or wholly discoverable from the covariates.

To account for covariates in ASP linkage analysis, we propose two mixture models, the pre-clustering and cov-IBD models. The basic form of the models is identi-

TABLE II. Type-I Error Rate of Pre-clustering and Cov-IBD Tests Based on 10,000 Simulations of Sib-Pair Samples of Size $n = 100$

Age at onset[a]	Method	Median *P*-value	Median λ	Nominal significance level		
				level = 0.05	level = 0.01	level = 0.001
35	Pre-cluster	0.5	1	0.0510	0.0093	0.0009
	Cov-IBD	0.5	1	0.0496	0.0081	0.0011
65	Pre-cluster	0.5	1	0.0513	0.0115	0.0014
	Cov-IBD	0.5	1	0.0488	0.0092	0.0004

[a]Maximum age at onset of proband.

cal (Eq. 1), but they differ in some key features. In the pre-clustering model, weights of membership in the linked and unlinked groups are determined prior to any linkage analysis. Membership information, therefore, comes only from the values of the covariates for ASP and, possibly, auxiliary information about those covariates. Any form of probabilistic clustering can be used to derive the weights. Auxiliary biological information can be invaluable because it will determine which group is of interest (i.e., potentially linked). We propose a likelihood ratio test for linkage for the pre-clustering model, in which the prior probability that a family is in the linked subset is estimated using a clustering model on the covariates. Adapting the results of Liang and Rathouz [1999], we show the test has a simple, limiting distribution. Moreover, it is simple to implement, making it a natural approach for genome scans, and gives a substantial increase in power when compared with a likelihood-ratio test that does not utilize covariate information.

The cov-IBD model is not as simple but is potentially more powerful because it fully utilizes all the information in the model; however, it does so at a cost in DF. In this model, the probability of membership in the linked and unlinked groups is again determined by the ASP covariate values, but, unlike the pre-clustering model, IBD values also contribute information for group membership. Group membership is taken to be a latent variable in a logistic model, with the parameters of the logistic model estimated indirectly from the full likelihood. In practice, this method has three drawbacks: 1) the likelihood ratio test does not have a simple approximate limiting distribution for tail probabilities; 2) the likelihood surface may possess multiple modes, and without a careful choice of starting values, a standard optimization routine may fail to find the global maximum; and 3) an arbitrary lower bound on the missing weights is required and needs some care to implement correctly. Chen et al. [2001] offer an alternative solution that also attempts to circumvent the difficulties encountered when conducting a likelihood-ratio test for mixture models. Their solution involves utilizing a penalized likelihood approach. Although the penalized likelihood appears to be simpler to implement, it does not naturally incorporate constraints that force the desired cluster to be the linked group. For this reason, we do not investigate this approach in this report.

Because ASP typically do not convey substantial IBD information to discriminate between linked and unlinked groups, we do not expect cov-IBD to perform substantially better than the pre-clustering model: for multiple covariates, in fact, we expect it to perform worse. On the other hand, for family structures with far greater IBD information, such as families with more than two affected siblings, the cov-IBD method should perform much better. We currently are exploring the performance of cov-IBD for nuclear families with more than two affected individuals.

It is permissible to investigate a large number of covariates, independent of IBD information, to detect clustering without incurring a penalty in the size of the test for linkage. It is only when the linkage analysis itself is performed multiple times that penalties for multiple testing must be imposed. For instance, it is permissible to perform cluster analysis with many different covariates to identify those covariates that potentially separate the data into homogeneous groups. However, if the linkage analysis is performed many times with either method, then a Bonferroni correction is appropriate. For this reason, we recommend that users investigate the covariates graphically using the clustering methods described for the pre-clustering

model, even if ultimately the analysis will be conducted using the cov-IBD model. Otherwise, the analysis may ultimately lose power [Leal and Ott, 2000].

Schaid et al. [2001] also developed a mixture model for linkage analysis with covariates. Their parametric linkage model has similarities to our models, especially cov-IBD. The models are not identical, however, and are in some ways complementary: Schaid et al. focus on a test of whether covariates discriminate between linked and unlinked pedigrees, given linkage, whereas our models test for linkage allowing for heterogeneity, which is determined, in part or in total, by covariate values.

Several other methods of ASP linkage analysis with covariates are available. The most recent work traces to Rice [Dorr et al., 1997; Rice et al., 1999], Greenwood and Bull [1997, 1999a], and Olson [1999]. Greenwood and Bull [1997, 1999a] and Olson [1999] formulate general conditional multinomial logistic models useful even when IBD information is incomplete. Rice et al. [1999; see also Dorr et al., 1997] use transmission of alleles from parent to ASP to set up a natural logistic model. For continuous covariates, the underlying assumption of all these models is that IBD status varies continuously with covariate values for the ASP. This assumption will be a good approximation to nature in some cases, but it will be hard to know a priori what those cases are. For a single, binary covariate, such as exposure status, all these models share features with the cov-IBD model.

The method of ordered-subsets analysis [Hauser et al., 1998] takes a less formal approach to incorporating covariates into the analysis. With this approach, families are ranked according to their mean sibship value for the quantitative trait of interest. Starting with the family at the extreme value for the covariate, the maximum lod score is computed for all ordered subsets of families. The final result is the maximum lod score for the subset yielding the maximum value of the test statistic over all subsets. To account for multiple testing, a permutation test is performed. This approach is similar in spirit to both methods presented in this article.

In contrast to ordered subsets, pre-clustering has the advantage of allowing for a specific comparison and therefore greater power if that group is well chosen. It also has the advantage of a known asymptotic distribution for the test statistic. Ordered subsets have the advantage of facilitated exploration of variables in a data-driven, hypothesis-generating framework, in the situation in which a priori there is not one or even a few obvious clusterings on which to base an analysis. Cov-IBD has the advantage of allowing for multiple covariates simultaneously, while ordered subsets can quickly and easily be carried out without worry about convergence and boundary issues and can be used on general pedigrees without the need to break families into relative pairs. Further, ordered subsets can be used easily with any additive linkage statistic.

A major strength of the mixture models that we describe, at least for some covariates, is that they approximately model the process generating the data by focusing on a linked subset. Another strength of these models is that they make explicit what attributes of the data are of interest. For these mixture models, researchers who are trying to discover the genetic basis of complex disease should be looking for distinct groups in the data, as determined by the covariates.

Olson [1999] extends the conditional logistic framework to account for ARP as well as ASP by assuming Risch's [1990b] likelihood model for allele sharing. Our mixture models also can be extended to ARP following Olson's lead. In fact, the

generalization is mathematically straightforward, involving substitution of the null and alternative ASP allele-sharing distributions with those appropriate for particular ARP. The biological meaning of the extension, however, may not be so simple. In particular, for ARP we would be cautious about covariates that are biological traits because the trait realizations could be affected by age, period, or cohort differences, which may not be easily captured by clustering models. The same problem arises for the logistic models.

Our methods have been developed under the assumption of a single ASP per family. When a few families in the sample have more than two affected siblings, which is a common situation, it seems sensible to break these families into all possible ASP and treat the pairs as if they were independent. For the models that we present, this treatment is only slightly biased [Greenwood and Bull, 1999b], so the true *P*-values will be very close to their nominal values. When a large fraction of the families has more than two affected siblings, a different approach is desirable. A more ambitious extension to our model, which we are pursuing, allows for families with arbitrary pedigrees using a nonparametric likelihood. Clearly IBD sharing within larger families will be more informative in determining the linkage status of families than IBD status from a single relative pair. We anticipate that in this setting the cov-IBD model is likely to far outperform the pre-clustering method.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health grants MH57881 and MH56193 and NSF grant DMS 9803433. The authors are grateful to Dan Weeks for helpful comments on this manuscript and to the University of Pittsburgh Statistical Genetics Journal Club for stimulating discussion of the topic. The program SibCov, which performs all the analyses described in this manuscript, can be obtained at http://wpicr.wpic.pitt.edu/WPICCompGen/ or http://www.stat.cmu.edu/~roeder/. SIMULATE is available at ftp://linkage.rockefeller.edu/software/simulate/.

APPENDIX: PRE-CLUSTERING ASYMPTOTICS

We prove that $\tilde{\Lambda} \sim \chi^2_1$ when $\tilde{\lambda}$ is unrestricted; then the mixture of χ^2's result is obtained as a direct application of Self and Liang [1987]. Throughout we use $\dot{f}$ to mean the derivative of f with respect to λ. After plugging in $\tilde{\alpha}(z_i)$ for α, the pseudo-likelihood is of the form $\log L(\lambda) = \Sigma_i \log[\tilde{\alpha}(z_i)g(x_i; \lambda) + (1 - \tilde{\alpha}(z_i))g_0(x_i)]$. The score for λ equals $U(\lambda) = \Sigma_i \tilde{\alpha}(z_i)\dot{g}(x_i; \lambda)/p[x_i; \lambda, \tilde{\alpha}(z_i)]$. We require that $U(\lambda)$ is an unbiased estimating equation for λ when $\lambda = 1$, but this follows directly by interchanging the order of differentiation and integration: $E[U(1)] = 0$. The variance of the score at $\lambda = 1$ is approximately equal to $n\sigma^2$, where σ^2 is defined as

$$\sigma^2 = \frac{1}{n}\sum_i \tilde{\alpha}(z_i)^2\{\dot{g}(x_i;1)/g_0(x_i)\}^2.$$

By the central limit theorem, $n^{-1/2}U(1) \to N(0, \sigma^2)$. We define $\tilde{\lambda}$ as the quantity obtained from maximizing equation (1), where $\tilde{\alpha}(z_i)$ replaces $\alpha(z_i)$ in the likelihood. For the proof, we approximate $\tilde{\lambda}$ by the one-step estimator. Taking a Taylor series

expansion of $U(1)$ at $\tilde{\lambda}$ we get $U(1) \approx U(\tilde{\lambda}) - (\tilde{\lambda} - 1)\dot{U}(1)$. Note that $U(\tilde{\lambda}) = 0$ by definition of the maximum likelihood estimate. Because the variance of the score equals the expected value of the negative derivate of the score, we obtain $-\dot{U}(1) = n\sigma^2$. We can conclude $n^{1/2}(\tilde{\lambda} - 1) \approx n^{-1/2}U(1)/\sigma^2$. Finally, we expand $\tilde{\Lambda}$ about $\lambda = 1$ to obtain the asymptotic distribution of the score statistic $\tilde{\Lambda} \approx [n^{1/2}(\tilde{\lambda} - 1)]^2\ \sigma^2$. It follows that when $\tilde{\lambda}$ is restricted to ≥ 1, $\tilde{\Lambda}$ converges in distribution to a one-sided χ^2_1.

REFERENCES

Banfield JD, Raftery AE. 1993. Model-based Gaussian and non-Gaussian clustering. Biometrics 49:803–22.

Chen H, Chen J, Kalbfleisch JD. 2001. A modified likelihood ratio test for homogeneity in finite mixture models. J R Stat Soc B 63:19–29.

Chakravarti A, Badner JA, Li CC. 1987. Tests of linkage and heterogeneity in Mendelian diseases using identity by descent scores. Genet Epidemiol 4:255–66.

Chernoff H, Lander E. 1995. A symptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomical. J Stat Plan Inf 43:19–40.

Claus EB, Risch NJ, Thompson WD. 1990. Using age of onset to distinguish between subforms of breast cancer. Ann Hum Genet 54:169–177.

Dawson DV, Kaplan EB, Elston RC. 1990. Extensions to sib-pair linkage tests applicable to disorders characterized by delayed onset. Genet Epidemiol 7:453–466.

Dorr DA, Rice JP, Armstrong C, Reich T, Blehar M. 1997. A meta-analysis of chromosome 18 linkage data for bipolar illness. Genet Epidemiol 14:617–22.

Finch SJ, Mendell NR, Thode HC. 1989. Probabilistic measures of adequacy of the numerical search for a global maximum. J Am Stat Assoc 84:1020–23.

Fraley C, Raftery AE. 1998. MCLUST: Software for Model-Based Cluster and Discriminant Analysis Technical Report No. 342, University of Washington (http://www.stat.washington.edu/fraley/mclust/reps.shtml).

Ghost JK, Sen PK. 1985. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In: Le Cam LM, Olshen RA, editors. Proceedings of the Berkeley Conference in Honor of J. Neyman and J. Kiefer, Vol. II. Belmont, CA: Wadsworth. pp 789–806.

Ghosh S, Watanabe RM, Valle TT, Hauser ER, Magnuson VL, Langefeld CD, Ally DS, et al. 2000. The Finland-United States Investigation of NonInsulin-Dependent Diabetes Mellitus Genetics (FUSION) Study. I. An Autosomal Genome Scan for Genes That Predispose to Type 2 Diabetes. Am J Hum Genet 67:1174–85.

Goddard KAB, 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 68:1197–206.

Greenwood CM, Bull SB. 1997. Incorporation of covariates into genome scanning using sib-pair analysis in bipolar affective disorder. Genet Epidemiol 14:635–40.

Greenwood CM, Bull SB. 1999a. Analysis of affected sib pairs, with covariates—with and without constraints. Am J Hum Genet 64:871–85.

Greenwood CM, Bull SB. 1999b. Down-weighting of multiple affected sib pairs leads to biased likelihood-ratio tests, under the assumption of no linkage. Am J Hum Genet 64:1248–52.

Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King M-C. 1990. Linkage of early-onset familial breast cancer to chromosome 17q21. Science 250:1684–9.

Hauser ER, Watanabe RM, Duren WL, Boehnke M. 1998. Stratified linkage analysis of complex genetic traits using related covariates. Am J Hum Genet Suppl 63:A45.

Holmans P. 1993. Asymptotic properties of affected-sib-pair linkage analysis. Am J Hum Genet 52:362–74.

Iversen Jr ES, Parmigiani G, Berry D, Schildkraut J. 2000. Genetic susceptibility and survival: application to breast cancer. J Am Stat Assoc 95:28–41.

Jones B, Nagin D, Roeder K. 2001. A SAS Procedure based on Mixed Models for Estimating Developmental Trajectories. Sociol Methods Res 29:374–93.

Jorgenson E, Hinds D, Risch N. 1999. Sib-pair analysis of the collaborative study on the genetics of alcoholism data set. Genet Epidemiol 17(suppl):S187–91.

Kaufman L, Rousseeuw PJ. 1990. Finding groups in data: an introduction to cluster analysis. New York: John Wiley & Sons.

Kong A, Cox NJ. 1997. Allele-sharing models: LOD scores and accurate linkage tests. Am J Hum Genet 61:1179–88.

Kruglyak L, Daly MJ, Reeve-Daly MJ, Lander ES. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347–63.

Leal SM, Ott J. 2000. Effects of stratification in the analysis of affected-sib-pair data: benefits and costs. Am J Hum Genet 66:567–75.

Lemdoni M, Pons O. 1995. Tests for genetic linkage and homogeneity. Biometrics 51:1033–41.

Liang K-Y, Rathouz PJ. 1999. Hypothesis testing under mixture models: application to genetic linkage analysis. Biometrics 55:65–74.

Matise TC, Weeks DE. 1993. Detecting heterogeneity with the affected-pedigree-member (APM) method. Genet Epidemiol 10:401–6.

Olson JM. 1999. A general conditional-logistic model for affected-relative-pair linkage studies. Am J Hum Genet 65:1760–69.

Ott J. 1983. Linkage analysis and family classification under heterogeneity. Ann Hum Genet 47:311–20.

Rice JP, Rochberg N, Neuman RJ, Saccone NL, Liu KY, Zhang X, Culverhouse R. 1999. Covariates in linkage analysis. Genet Epidemiol 17(suppl):S691–5.

Risch N. 1990a. Linkage strategies for genetically complex traits. I. Multilocus models. Am J Hum Genet 46:222–8.

Risch N. 1990b. Linkage strategies for genetically complex traits. II. The power of affected relative pairs. Am J Hum Genet 46:229–41.

Risch N. 1990c. Linkage strategies for genetically complex traits. III. The effect of marker polymorphism on analysis of affected sib pairs. Am J Hum Genet 46:242–53.

Roeder K, Lynch K, Nagin D. 1999. Modeling uncertainty in latent class membership: a case study in criminology. J Am Stat Assoc 94:766–76.

Schaid DJ, McDonnell SK, Thibodeau SN. 2001. Regression models for linkage heterogeneity applied to familial prostate cancer. Am J Hum Genet 68:1189–96.

Self SG, Liang K-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. J Am Stat Assoc 82:605–10.

Smith CAB. 1963. Testing for heterogeneity of recombination fraction values in human genetics. Ann Hum Genet 27:175–82.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–16.