A Unified Approach to Modelling Linkage to Quantitative and Qualitative Traits

J. J. Houwing-Duistermaat¹, C. Bijkerk¹, L. Hsu², Th. Stijnen¹, E. P. Slagboom³ and C. M. van Duijn¹
¹Department of Epidemiology and Biostatistics, Erasmus Medical Center, Rotterdam, The Netherlands ²Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA, ³Department of Medical Statistics, Leiden University Medical Center, Leiden, The Netherlands

Summary

For quantitative traits with a genetic component, random effects approaches are used to test for linkage at observed marker loci. We propose to use these approaches also for binary outcomes observed in sib pairs derived from a population-based cohort study. In addition to a random effect modelling correlation due to polygenic effect, a random effect is included to model the correlation between siblings due to sharing alleles identical by descent (IBD) at the observed marker locus. A two-step analysis is proposed. Firstly, score statistics are computed to test whether correlation is present in the data. Secondly, random effects models are fitted, yielding heritability estimates. To illustrate the methods, data on the contribution of the COL2A1 gene to various binary and quantitative outcomes including the presence of Heberden’s nodes and bone mineral density (BMD) are analysed. For most of the traits studied, the score statistics were significant, indicating the presence of genetic effects. For BMD and for Heberden’s nodes, the variance explained by the marker locus was 44% (P = 0.0008) and 15% (P = 0.38) respectively. We conclude that the score statistics can be used as a preliminary data analysis. In more sophisticated analysis, heritabilities can be estimated by fitting random effects models.

Keywords: sibships, population, score test, complex genetic traits, random effects

Introduction

In the past, many genes have been identified that contribute to monogenic Mendelian diseases. For these diseases, the underlying genetic model was known and could be specified in a classical linkage analysis (Ott, 1999). Locating genes is more complicated if traits are studied that are influenced by a large number of environmental and genetic factors. For these traits, the genetic parameters such as number of genes involved, allele frequencies and penetrances are often unknown. Therefore allele-sharing methods applied to sib pair data which do not assume a genetic model are popular tools to map these traits (Olson et al. 1999).

For quantitative traits, a generalized estimating equation approach (GEE) has been used to estimate variances of random effects which represent a quantitative trait locus (Haseman & Elston, 1978; Elston et al. 2000; Amos et al. 2000; Sham & Purcell, 2001). The advantages of GEE methods are that it is not necessary to fully specify the distribution of the random effects. An alternative approach is to specify the full distribution of the data and to maximize the loglikelihood function. Several authors used this approach to estimate the variances of normally distributed random effects which model the correlation caused by IBD sharing of marker alleles (Amos, 1994; Amos et al. 1996; Fulk et al. 1999; Williams & Blangero, 1999). Especially for sibships of size larger than two, the GEE based estimators may be less efficient than the maximum likelihood estimator because not all available information is used (Liang & Zeger, 1986; Amos et al. 1996).

© University College London 2003

Also for binary outcomes, we consider studies with data available on randomly obtained sibships. For pedigree analysis, Williams et al. (1999) included random effects using probit link to model linkage, fitting the model by approximating the log likelihood. Previously, we considered a random effect approach to model several causes of correlation in families (Houwing-Duistermaat et al. 1998; 2000). A logit link function was used and the random effects were assumed to follow a normal distribution. The variance components were estimated using a pseudo profile likelihood approach. In this paper, the same approach is proposed to estimate variances of random effects modelling allele sharing at marker loci for binary outcomes observed in sib pairs derived from a population-based sample.

Before fitting the models to binary or quantitative outcomes, we propose to apply a score statistic to test for the presence of genetic correlation. This test does not assume a normal distribution for the random effects. In a genome wide scan, the test statistic can be used to identify interesting regions, hence saving a lot of time for the researcher by limiting the number of random effect models to be fitted to marker loci in candidate regions. We propose to compute these score statistics as a first step in linkage analysis. In the second step, the random effect models can be estimated using maximum likelihood methods.

In this paper, the methods are described for one typed marker. They are applied to an empirical study on the relation between the candidate gene COL2A1 and osteoarthritis, Heberden’s nodes and bone mineral density (BMD). Linkage is modelled between a trait locus linked to a VNTR marker next to the COL2A1 gene and the various outcomes observed in sibship clusters.

Methods

Let \( Y_i \) be a response vector of members of randomly chosen sibship \( i \) and \( Y_{ij} \) the response variable of sibling \( j \) of sibship \( i \). The outcomes of sibships may be correlated due to genes involved in the etiology of the observed trait. Let \( u_i^g \) be the vector of these genetic effects for sibship \( i \) and let \( u_{ij}^g \) be the genetic effect of sibling \( j \). This genetic effect has zero-mean, variance \( \tau_g^2 \) and correlation structure \( G_i \), where the \( jk \)th element of \( G_i \) is equal to a half, because on average siblings share half of their genes. Under the additional assumption that \( u_i^g \) follows a normal distribution, \( u_i^g \) is an additive polygenic effect.

Next, we consider the information available at a marker locus. Outcomes of siblings may be correlated due to a genetic factor tightly linked to the observed marker locus. In the case of a marker located next to a candidate gene (as in our data example) or a genome wide scan, the identical by descent (IBD) status at the marker locus can be assumed to be equal to the IBD status at the trait locus. For simplicity, we assume an additive genetic model for this single gene. Let \( u_{ij}^g \) be the vector of these marker effects for sibship \( i \). It has zero-mean, variance \( \tau_{jk}^2 \) and correlation structure \( S_i \), where the \( jk \)th element of \( S_i \) is equal to the proportion \( \pi_{i,jk} \) of shared marker alleles IBD between sibling \( j \) and sibling \( k \) of sibship \( i \). When the number of shared alleles is not observed, \( \pi_{i,jk} \) is estimated by \( \hat{\pi}_{i,jk} = 0.5 \hat{f}_{ij} + \hat{f}_{2ij/k} \) with \( \hat{f}_{ij} \) the estimate of the probability that sibling \( j \) and sibling \( k \) share \( l \) alleles IBD. For randomly selected sibships, the same results are obtained by using \( \hat{\pi}_{i,jk} \) instead of using the full distribution of IBD sharing (Fulker & Cherny, 1996; Dolan et al. 1999).

These two genetic effects \( u_i^g \) and \( u_{ij}^g \) are random effects and the following family of random effect models is proposed to model the data:

\[
E(Y_i | u_i^g, u_{ij}^g) = h^{-1}(X_i^\beta + u_i^g + u_{ij}^g), \tag{1}
\]

where \( h \) is a link function (McCullagh & Nelder, 1989), \( X_i \) is the matrix of covariates and \( \beta \) are the regression coefficients. For binary outcomes, the logit function is used and for quantitative outcomes, the identity is used as link function. It is assumed that \( Y_i^g \) are conditionally independent given the genetic effects \( u_i^g \) and \( u_{ij}^g \). The heritability \( \hat{h}_i^2 = \frac{\hat{\tau}_g^2 + \hat{\tau}_{jk}^2}{\hat{\tau}_g^2 + \hat{\tau}_{jk}^2 + \hat{\sigma}_i^2} \) measures the total effect of genetic factors on a trait and the locus specific heritability \( \hat{h}_i^2 = \frac{\hat{\tau}_{jk}^2}{\hat{\tau}_g^2 + \hat{\tau}_{jk}^2 + \hat{\sigma}_i^2} \) measures the contribution of a locus to the trait. For quantitative traits, \( \hat{\sigma}_i^2 \) is equal the residual variance of \( Y \) and under the standard logistic distribution for qualitative traits, \( \hat{\sigma}_i^2 \approx 3 \) (Bickel & Docksum, 1977). The variance of \( Y_i^g \) given \( u_i^g \) and \( u_{ij}^g \) represents the unpredictability of the response variable given the genetic effects. Note that by adjusting for covariates, we hope to minimise the correlation due to shared environmental influences.
factors may still explain a part of the correlation within sibships.

We propose a two-step analysis. The first step is to test the for the presence of a prespecified correlation structure using a score statistic (Houwing-Duistermaat et al. 1995; 1998). These score tests are robust against deviations of the normal distribution of the random effects, time efficient and adjustments are made for observed covariates. The correlation of $u_i^2 + u_e^2$ in model (1) equals $R = \rho S + (1 - \rho) G$ with $\rho = \frac{\tau_1^2}{\tau_1^2 + \tau_2^2}$ unknown. For a specific $R$, the test statistic $Q^R$ is

$$Q^R = \frac{\sum_{i=1}^n (Y_i - \hat{\mu}_i)'R(Y_i - \hat{\mu}_i)}{\sum_{i=1}^n (Y_i - \hat{\mu}_i)'(Y_i - \hat{\mu}_i)}$$

(2)

Here, $Y_i$ and $\hat{\mu}_i$ are vectors of outcomes and $h^{-1}(X_i\hat{\beta})$ respectively and $\hat{\beta}$ is estimated by maximizing the likelihood under the null hypothesis of no correlation ($\tau_1^2 = \tau_2^2 = 0$). The sums are taken over the number of observed sibships $n$. The distribution of $Q^R$ may be approximated with a scaled chi-square distribution (for details see Houwing-Duistermaat et al. 1995). We propose to test for two values of $\rho$, namely $\rho = 0 (R = G)$ and $\rho = 1 (R = S)$. Adjustments for multiple testing may be made by using a standard correction. Alternatively one may be interested in the maximum testing may be made by using a standard correction. Alternatively one may be interested in the maximum likelihood value $l_p$ in $(\hat{\beta}(\tau_1^2, \tau_2^2), \tau_1^2, \tau_2^2)$ is estimated by using Markov chain Monte Carlo to integrate over the distributions of the random effects $u_i$ and $u_e$. Under regularity conditions the estimators $\hat{\tau}_1^2$ and $\hat{\tau}_2^2$ equal to those values for which the log likelihood surface $l_p$ has its maximum value, are consistent.

The null hypothesis of no linkage at a marker locus while taking into account the presence of unlinked genetic factors may be tested by fitting a random effect model with only the genetic effect $u_i^2$ and using a likelihood ratio statistic to compare this model with model (1). Two times this difference in log likelihood follows a fifty–fifty mixture of a chi-square distributions of 0 degrees of freedom and 1 degree of freedom.

Application

Data

The methods are applied to a population based study on the contribution of the candidate gene COL2A1 to radiological osteoarthritis (ROA), Heberden’s nodes and bone mineral density of the femoral neck (BMD). Osteoarthritis (OA) is the most common disease of the musculoskeletal system (Van Saase et al. 1989). Radiographic changes in OA reflect a progressive deterioration of articular cartilage. Age, sex, body mass index (BMI) and BMD are determinants of ROA (Craemer & Hochberg, 1997). Heberden’s nodes are bulbous deformities at the distal interphalangeal joints of the fingers and are associated with OA in the hand joints (Kellgren & Moore, 1952). The COL2A1 gene (12q13) encoding
the cartilage protein, collagen type II, is a candidate gene for OA (Eyre et al. 1991; Kuivaniemi et al. 1991). The Vitamin D Receptor (VDR) gene is very close located to the COL2A1 gene and is a candidate gene for BMD as well as for OA (Uitterlinden et al. 1997). Hence statistical evidence for linkage may be explained by the candidate gene VDR as well as the COL2A1 gene.

The study is embedded in the Rotterdam Study; a prospective population based follow-up study of determinants and prognosis of chronic disease in about 8000 subjects aged 55 years and over (Hofman et al. 1991). In this population more than 95% of the subjects have two Caucasian parents. Data are available on 819 individuals aged 55-70 years randomly selected of the total data-set of the Rotterdam Study. From a selected subset of 112 probands, data are collected on their siblings (196 in total). These probands have been selected on the presence of ROA in two or more out of four joint sites (knee, hip, hand and spine) and are between 55 and 65 years of age. The sizes and percentages (between brackets) of the observed sibships, including the probands, are 1 (13%), 2 (48%), 3 (18%), 4 (8%), 5 (5%), 6 (1%), 7 (4%) and 8 (7%).

Note that we do not have information on the siblings of 707 probands. These probands are healthy or have only one affected joint. Since the ascertainment of siblings of persons is conditional on the outcome of probands, these data are missing at random (Diggle et al. 1994). Thus, the healthy persons and patients with only one affected side are included in the analysis as sibships of size 1. In this way estimates of population based variance components are obtained when maximum likelihood methods are used.

In this paper, the following binary outcomes are analyzed: ROA in knee, disk degeneration at spine and Heberden’s nodes (1 present, 0 not present). For ROA in hands, the sum is taken over the 16 joint groups studied (see Bijkerk et al. 1999). To study generalized osteoarthritis (GOA), the sum over all joint groups studied is computed. The maximum score is equal to 23. The logarithm of these sumscores and the trait BMD are analyzed as a quantitative traits.

The population cohort is described in table 1.

### Table 1 Baseline measurements consisting of observed means with standard deviations between brackets in the random population cohort of 818 individuals

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Males</th>
<th>Females</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 331 (40%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n = 487 (60%)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cohort n = 818</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Quantitative outcomes</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GOA</td>
<td>0.98</td>
<td>1.19</td>
<td>1.10</td>
</tr>
<tr>
<td>ROA at hand</td>
<td>0.47</td>
<td>0.80</td>
<td>0.66</td>
</tr>
<tr>
<td>Binary outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BMD (g/cm²)</td>
<td>0.91</td>
<td>0.85</td>
<td>0.87</td>
</tr>
<tr>
<td>Disk degeneration at spine</td>
<td>0.63</td>
<td>0.60</td>
<td>0.61</td>
</tr>
<tr>
<td>ROA at knee</td>
<td>0.13</td>
<td>0.22</td>
<td>0.18</td>
</tr>
<tr>
<td>Heberden’s nodes</td>
<td>0.16</td>
<td>0.28</td>
<td>0.23</td>
</tr>
</tbody>
</table>

* log transformed values.

For the analyses of the OA data, first the proportion of VNTR alleles shared IBD are estimated for each sib pair using the computer program Mapmaker\Sibs (Kruglyak & Lander, 1995). Then the score tests can be applied. In table 2, the p-values of the $Q^C$ and $Q^S$ statistics are given. For all quantitative traits, both test statistics give highly significant p-values. For Heberden’s nodes both age and sex. Since no other important environmental risk factors are known, it is likely that most of the correlation of the outcomes is caused by genetic factors.

Except for one proband, all the individuals (818 random individuals + 196 siblings) are typed for the VNTR polymorphism located 1.35 kilobases from the 3’ end of the COL2A1 gene, between the COL2A1 and the VDR gene. In the population based cohort of 818 individuals, 15 different alleles are present. In this cohort, the frequencies of the most common alleles are 0.42, 0.27, 0.11, 0.06, 0.06. The other allele frequencies are smaller than 0.05. The genotype frequencies are in Hardy Weinberg proportions ($P = 0.91$, using the package GENEPOP, Raymond & Rousset, 1995), indicating that we have indeed a random sample and that we can use these allele frequencies to compute the probabilities that a sib pair shares 1 or 2 alleles IBD $(f_1, f_2)$ when the parental genotypes are unknown.

### Analysis

For the analyses of the OA data, first the proportion of VNTR alleles shared IBD are estimated for each sib pair using the computer program Mapmaker\Sibs (Kruglyak & Lander, 1995). Then the score tests can be applied. In table 2, the p-values of the $Q^C$ and $Q^S$ statistics are given. For all quantitative traits, both test statistics give highly significant p-values. For Heberden’s nodes both
Table 2 P-values of the statistics \( Q_G \) and \( Q_S \) in the total cohort of 818 randomly chosen individuals and 196 siblings of 112 selected probands

<table>
<thead>
<tr>
<th>Outcome</th>
<th>( Q_G )</th>
<th>( Q_S )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GOA(^*)</td>
<td>( 9.0 \times 10^{-6} )</td>
<td>( 2.5 \times 10^{-5} )</td>
</tr>
<tr>
<td>ROA at hand(^*)</td>
<td>( 1.7 \times 10^{-8} )</td>
<td>( 2.8 \times 10^{-7} )</td>
</tr>
<tr>
<td>BMD(^+)</td>
<td>&lt; 1 \times 10^{-8}</td>
<td>&lt; 1 \times 10^{-8}</td>
</tr>
<tr>
<td>Binary outcomes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Disk degeneration at spine(^*)</td>
<td>0.01</td>
<td>0.08</td>
</tr>
<tr>
<td>ROA at knee(^*)</td>
<td>0.90</td>
<td>0.93</td>
</tr>
<tr>
<td>Heberden’s nodes(^+)</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>

\(^*\) adjusted for age, sex, BMI and BMD.
\(^+\) adjusted for age and sex.
\(^\parallel\) log transformed values.

Discussion

For linkage studies on outcomes observed in sibship clusters embedded in a population based study, we propose a two-step random-effect approach. This approach gives a unified framework for both quantitative and binary outcomes. To reduce variability and to reduce the correlation due to other factors than genes involved in the aetiology of the trait, the outcomes can be adjusted for observed covariates. In the first step, score tests \( Q_R \) are applied which are robust. In the second step, the random effect models are fitted by maximizing the (pseudo) log likelihood. Advantages of the likelihood methods are that it is efficient and data missing-at-random are allowed. A drawback of these methods is that an assumption about the distribution of the genetic effects has to be made (Allison et al. 1999). Although a
Gaussian distribution may be unrealistic for rare causal variants, in most situations it yields a valid approximation. However, no assumption about the distribution of the random effects is made in the score tests $Q^s$.

The methods are applied to a study of the effect of sharing alleles IBD at a VNTR marker located between the COL2A1 gene and the VDR gene on various outcomes. For all quantitative traits, both correlation structures were highly significant. Only for BMD, adding a marker effect to a model with only an additive environmental effect improved the model significantly ($P = 0.0008$). The locus explained 44% of the total variance. For the binary outcome, Heberden nodes, the locus specific heritability was 15% ($P = 0.38$).

Especially for mapping complex common traits, we feel that it will be important to report population-based estimates of heritabilities in addition to $p$-values. However often only nonrandom sibship data are available. To apply the methods proposed in this paper to these data, the likelihood function has to be adjusted (see for example Sham et al. 2000) A corresponding score statistic could be derived from this likelihood function.

Recently it is shown that for randomly selected sib pairs, the weighted Haseman Elston approach (Sham & Purcell, 2001) is equivalent to the score statistic for testing the null hypothesis $\tau^2 = 0$ (Wang & Huang, 2002; Putter et al. 2002) while allowing for sibling correlation ($\tau^2 > 0$). For qualitative traits, it is not straightforward to derive this score statistic, due to the non linearity of the relation between outcome and random effect.

Our results suggest that for mapping of complex genetic traits, the random effects approach is a promising tool that also can be used for binary data. The approach comprises testing for the presence of genetic effects using the statistics $Q^s$ and estimation of the locus specific heritability.

### Acknowledgements

This study is supported by a Pioneer Grant of the Netherlands Organisation for Scientific Research (NWO). We would like to thank Professor Hans van Houwelingen for his comments and Dr. Ingrid Meulenbelt, TNO, Institute of Prevention and Health, Leiden, The Netherlands, for the coordination of the genotyping. SPLUS functions to compute the score statistics are available from http://www.geneticepi.com/Research.

### References


Modelling Linkage


Received: 18 February 2002
Accepted: 13 December 2002