Behav Genet (2010) 40:404–414
DOI 10.1007/s10519-009-9322-8

ORIGINAL RESEARCH

Matched Ascertainment of Informative Families for Complex Genetic Modelling

Benjamin H. Yip · Marie Reilly · Sven Cnattingius · Yudi Pawitan

Received: 18 September 2008 / Accepted: 28 November 2009 / Published online: 24 December 2009

Abstract Family data are used extensively in quantitative genetic studies to disentangle the genetic and environmental contributions to various diseases. Many family studies based their analysis on population-based registers containing a large number of individuals composed of small family units. For binary trait analyses, exact marginal likelihood is a common approach, but, due to the computational demand of the enormous data sets, it allows only a limited number of effects in the model. This makes it particularly difficult to perform joint estimation of variance components for a binary trait and the potential confounders. We have developed a data-reduction method of ascertaining informative families from population-based family registers. We propose a scheme where the ascertained families match the full cohort with respect to some relevant statistics, such as the risk to relatives of an affected individual. The ascertainment-adjusted analysis, which we implement using a pseudo-likelihood approach, is shown to be efficient relative to the analysis of the whole cohort and robust to mis-specification of the random effect distribution.

Keywords Segregation analysis · Mixed models · Variance components · Probit models

Edited by Stacey Cherny.

B. H. Yip
Department of Psychiatry, University of Hong Kong, Hong Kong, China

M. Reilly · S. Cnattingius · Y. Pawitan (✉)
Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 17177 Stockholm, Sweden
e-mail: yudi.pawitan@ki.se

Introduction

Family data have been used extensively for complex genetic modelling such as quantitative-trait linkage (e.g., Amos 1994; Blangero et al. 2001) or segregation analysis to separate genetic and environmental contributions to non-Mendelian diseases (e.g., Falconer 1965; Mather and Jinks 1977; Neale and Cardon 1992). Other than overcoming the sample size problem associated with twin studies, especially when the disease of interest has a low prevalence, family data potentially provide richer genetic information (e.g., Pawitan et al. 2004). However, this information is likely to be concentrated in 'genetically loaded' families, so that it is not efficient to collect data from, nor to analyse, all families from a population register. Non-random ascertainment is commonly used in genetics research to maximize the amount of information in the data for a given sample size (e.g., Elston and Sobel 1979). One of the most common methods of non-random ascertainment is to include families with at least one affected member: for variance component models this has been suggested, for example, in deAndrade and Amos (2000), Epstein et al. (2002), and Burton (2003). However, this sampling scheme may not be optimal, and in fact it has been shown (Glidden and Liang 2002; Noh et al. 2005) that the analysis of the ascertained data is sensitive to mis-specification of the random-effect distribution. In this paper we develop an efficient and robust method of ascertaining informative families from population-based family registers for the purpose of complex genetic modeling involving variance component analysis of a binary trait.

In epidemiological analyses, we often need or wish to account for confounding factors. For variance component analysis of a binary trait, the most straightforward way to adjust for potential confounders is to include them as

covariates in a generalized linear mixed model (GLMM) (Breslow and Clayton 1993; Lee and Nelder 1996). Marginal likelihood provides a flexible computational approach, and can be extended to multivariate binary traits analysis, as we have demonstrated in an analysis of the the co-morbidity of schizophrenia and bipolar disorder (Yip et al. 2008; Lichtenstein et al. 2009). The exact marginal likelihood computation is also more efficient than other computational approaches such as the Gibbs sampling (Zeger and Karim 1991; Burton et al. 1999), but for family data, it is still slow because of the high-dimensional integration (e.g., Pawitan et al. 2004). Because of this limitation, recent likelihood-based methods in family data analysis are limited in their ability to handle general covariates.

To avoid the integration step, Noh et al. (2006) used a hierarchical-likelihood method with Laplace approximation. However, for the volume of data that is typical for family studies, the computational requirements of these methods are still enormous, so they cannot be used during the model building stage, where numerous exploratory analyses are performed. Moger et al. (2008) suggested case-cohort methods as a way of dealing with large population-based family data with survival traits. We adapted their idea here and extended the exact marginal likelihood approach to a pseudo-likelihood approach to analyze ascertained family data.

Intuitively, information about familial clustering comes from families with at least two affected members. Thus, provided the genetic information in the full data can be preserved, ascertainment of families with at least two affected members offers the potential for dramatic data reduction; see Sect. 2.1 for a specific example. For computational efficiency it is natural to first group families with the same configuration of disease status and covariates. A novel aspect in our method is an ascertainment of the family configurations rather than family units. We propose an optimized matching method where we ascertain family configurations that are most informative, while making sure that relevant features, such as the risk to relatives of cases, are similar in the sampled data and the full cohort.

To summarize the contribution of this paper, we have developed a method to facilitate routine exploratory analysis of large population-based family data sets, where interest is focused on estimating the genetic and environmental contributions to a binary trait with adjustment for confounding. We propose an ascertainment scheme, where families are first grouped by the pattern of the outcomes and covariates of their members, and the ascertainment is of family configurations rather than family units. In our application, all families with two affected members are sampled, and the remaining families are sub-sampled in such a way that the sampled data matches the whole cohort with respect to the odds ratio for affected siblings. Our ascertainment-adjusted analysis, which uses a pseudo-likelihood method, is robust against mis-specification of the random-effect distribution and has high efficiency versus exact likelihood. We illustrate our method in a substantive analysis of a population-based dataset of birth outcome in pairs of siblings.

Methodology

SGA dataset

For motivation and illustration we use the small-for-gestational-age (SGA) data as described in Svensson et al. (2006). This dataset was obtained by linkage of the Swedish Multi-Generation Register and Medical Birth Register. We include covariates that have been suggested as potential risk factors to SGA, such as maternal age, preeclampsia diagnosis in an earlier pregnancy, smoking and body-mass index (BMI). Due to availability of information on some of the covariates, our data covers the calendar period (1981–2001). As in Svensson et al. (2006), we identified pairs of full siblings, where both of them had at least one delivery recorded in the Medical Birth Register, and we collected the birth information from the different types of sibships: sister–sister, brother–brother and sister–brother pairs.

The final dataset consists of 326,629 family-pairs (pairs of siblings with their spouses and offspring). The optimally matched sample is ascertained from these data, and its performance is compared to the analysis of the full data. To limit the data to a manageable volume, we used the information from a maximum of 4 pregnancies from any sib-pair. There were 129,593 family-pairs with 2 pregnancies, 125,405 with 3 pregnancies and 71,631 with 4 pregnancies. There were 921,925 offspring between 1981–2001, among whom we observed 21,103 born small for their gestational age (SGA). The following table shows the distribution of the number of SGA offspring within the families of sib-pairs:

Number of SGAs	0	1	2	3	4
Number of family pairs	306,706	18,807	1,055	58	3

Ascertaining only family-pairs with at least two affected members, we can limit the case family-pairs to just $1{,}055 + 58 + 3 = 1{,}116$, instead of $1{,}116 + 18{,}807 = 19{,}923$ if we consider at least one affected member.

It has been shown previously (Svensson et al. 2006) that genetic factors, especially the fetal component, account for the majority of the liability of having an SGA birth. However, birth order was the only fixed covariate included

in the model, while, as the authors pointed out, the genetic liability to SGA may be partly mediated by well-known maternal risk factors for SGA births, such as smoking and preeclampsia.

Data structure and likelihood

Let $y_i \equiv (y_{i1}, \ldots, y_{in_i})$ be the vector of binary outcomes from n_i members of family i, for $i = 1,\ldots, N$. The families are assumed to be independent. Let $x_1,\ldots, x_N$ be the corresponding covariate matrices, each of size $n_i \times p$. Also available is the information on relationships between members of a family, thus determining structures such as full siblings, cousins, paternal-halfsibs, etc. Conditional on the random effect b_i, we assume y_{ij} to be an independent Bernoulli with parameter p_{ij}, following a general linear mixed model (GLMM)

$$g(p_{ij}) = x'_{ij}\beta + z'_{ij}b_i,$$

where g() is a link funciton, β is a p-vector of fixed regression parameters. The random parameter b_i captures the dependencies between family members; the design vector z_{ij} shows the contribution of b_i to the outcome. To complete the specification, we assume b_i is normal with mean zero and variance $D_i(\theta)$, where θ contains all the variance component parameters.

In GLMM framework, the logit link is the canonical link function for binary-trait models. However, there are at least two reasons why the probit link may be preferred. Firstly, the probit link fits directly in the liability model (Sham 1998), which is commonly used in biometrical genetics applications. Secondly, the probit link also led to a convenient computation of the marginal likelihood in terms of multivariate normal probabilities (Pawitan et al. 2004). Noh et al. (2006) illustrated that the parameters estimated from the two models with different link functions are comparable after adjustment by a simple scale factor.

Specifically for the SGA data, a family structure consists of a pair of nuclear families made by full siblings. The vector y_i is the pregnancy outcomes from the two families. (The pregnancies are treated as the offspring of the families.) We consider the model (now in vector notation)

$$\Phi^{-1}(p_i) = x_i\beta + m_i + f_i + c_i + s_i, \quad (1)$$

where m_i is the vector of maternal effects, f_i the fetal effects, c_i the common couple environment effect and s_i the common sibling environment. The common couple environment is the unique environment created by the father and the mother, and the sibling environment is the common childhood and adolescent environment experienced by the siblings. The common family environment is the unique environment created by the father and mother, and the sibling environment is the common childhood environment experienced by the sisters. We assume that $m_i \sim N(0, \sigma_m^2 R_m)$, $f_i \sim N(0, \sigma_f^2 R_f)$, $c_i \sim N(0, \sigma_c^2 R_c)$ and $s_i \sim N(0, \sigma_s^2 R_s)$. To illustrate the discrepancy in the correlation matrices for the random effects, let assume a sister–sister pair family where each sibling had two preqnancies. The outcome y_i is a binary vector that indicates SGA status of the 4 pregnancies, and

$$R_m = \begin{pmatrix} 1 & 1 & 1/2 & 1/2 \\ 1 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1 \\ 1/2 & 1/2 & 1 & 1 \end{pmatrix},$$

$$R_f = \begin{pmatrix} 1 & 1/2 & 1/8 & 1/8 \\ 1/2 & 1 & 1/8 & 1/8 \\ 1/8 & 1/8 & 1 & 1/2 \\ 1/8 & 1/8 & 1/2 & 1 \end{pmatrix},$$

$$R_c = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \quad R_s = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

The (1,2)-element of R_m is equal to one since the first two outcomes come from the same mother, i.e. the first sister. The (1,2)-element of R_f is 0.5 since the first two foetuses are full siblings and the (1,3)-element of R_f is 0.125 since it refers to a cousin pair. Similar reasoning applies to R_c and R_s. For more details, we refer the reader to Pawitan et al. (2004).

From the probit model, we have

$$\begin{aligned} p_{ij} &= P(Z_j < x'_{ij}\beta + z'_{ij}b_i) \\ &= P(Z_j - z'_{ij}b_i < x'_{ij}\beta), \end{aligned}$$

where the Z_j's are independent standard normal variates. Thus we have the marginal probability

$$\begin{aligned} &p(Y_i = y_i|x_i) \\ &\quad = \int p(y_i|b_i)|D_i(\theta)|^{-q/2} \exp\{-\frac{1}{2}b'_i D_i(\theta)^{-1} b_i\} db_i \quad (2) \end{aligned}$$

$$= E_{b_i}\left\{\prod_j p_{ij}^{y_{ij}}(1-p_{ij})^{1-y_{ij}}\right\} \quad (3)$$

$$= P(l_{ij} < V_{ij} < u_{ij}, \text{for all} j), \quad (4)$$

where q is the dimension of b_i, and $V_{ij} \equiv Z_j - z'_{ij}b_i$. The vector $V_i \equiv (V_{i1},\ldots, V_{in\_i})$ is $N(0, \Sigma_i)$ with

$$\Sigma_i = z_i D_i(\theta) z'_i + I_i,$$

where z_i denotes the matrix obtained by stacking the row vectors z_{ij}, and I_i is the $n_i \times n_i$ identity matrix. The upper bound $u_{ij} = x'_{ij}\beta$ if $y_{ij} = 1$, and $u_{ij} = \infty$ if $y_{ij} = 0$. Similarly,

the lower bound $l_{ij} = -\infty$ if $y_{ij} = 1$, and $l_{ij} = x'_{ij}\beta$ if $y_{ij} = 0$. Computation of the normal probability (4) is done using a Monte–Carlo algorithm (Genz 1992).

Let $f_i(y_i, x_i, \beta, \theta) \equiv P(Y_i = y_i|x_i)$, where we make the parameters explicit; the total log-likelihood would then be

$$l = \sum_{i=1}^{N} \log f_i(y_i, x_i, \beta, \theta).$$

For the SGA data, N is of the order of 325,000 family-pairs. Since the evaluation of each probability requires a non-trivial Monte–Carlo integration, a naive approach is out of the question. Our problem is compounded by the fact that the resulting likelihood is not smooth, while we need to use a derivative-free optimization method.

The likelihood computation will obviously be faster if the data are grouped according to the configurations of the family outcomes and covariates $\{Y_i, x_i\}$. The total likelihood can then be written as

$$l = \sum_{k=1}^{M} w_k \log f_k(y_k, x_k, \beta, \theta), \tag{5}$$

where w_k is the number of families with the kth configuration, and M is the total number of configurations.

If the family data consist of information on binary outcomes and p binary covariates from k family members, then $M \leq 2^{k(p+1)}$, and the number of probability computations can be reduced by a factor of N/M. However, for analysis of families with up to 4 members, this grouping will substantially reduce the computation time when the we only use one or two covariates. As we increase the number of covariates, M increases rapidly, so even the grouped data become too large to analyse with exact methods. For the SGA data, with one covariate we have $M = 185$, but with 5 covariates it increases to more than 11,000.

Ascertainment

For the grouped data, ascertainment is naturally done on the family configurations rather than on the family units. Let $S = \{1,\ldots, M\}$ be the index set of all family configurations, and suppose that S can be divided into two disjoint sets, $S = S_0 \cup S_1$, where S_1 is the set of all families with at least k affected members, and S_0 is the set of control families. In line with the usual case–control studies, we will keep all case-family configurations. Control-family configurations will in general be included with probability less than one.

Exact and weighted likelihoods

Let $A_j = 1$ if family j is ascertained, and 0 otherwise, and $a_j = P(A_j = 1)$. Typically a_j is a function of the number of affected members, but it can also be a function of covariates. Then the exact ascertainment-adjusted likelihood contribution from an observed y_j is

$$P(Y_j = y_j|x_j, A_j = 1) = \frac{a_j P(Y_j = y_j|x_j)}{\sum_k a_k P(Y_k = y_k|x_j)},$$

where k runs over all possible configurations from the same covariate x_j, such that $\sum_k P(Y_k = y|x_j) = 1$. Note that the denominator needs the evaluation of probabilities for all families that might get ascertained, even if many of those are in fact unobserved. Thus the computational burden of the exact likelihood is still too demanding for routine analysis.

We instead consider a weighted-likelihood

$$\widehat{l} = \sum_{k=1}^{M} \frac{A_k}{P(A_k = 1)} w_k \log f_k(y_k, x_k, \beta, \theta), \tag{6}$$

which is clearly an unbiased estimate of the log-likelihood (5). The main advantage over the exact likelihood is that we only need to evaluate the probabilities for family configurations that are both observed and ascertained.

Computation and inference

Because of the Monte Carlo approximation, the log-likelihood (6) is not smooth. We use the derivative-free Nelder-Mead simplex algorithm (Nelder and Mead 1965) to get near to the solution, then use the Gauss-Seidel method with the smoothed log-likelihood to arrive at the final solution. The statistical software R was used for all computations.

Standard inference in the pseudo-likelihood framework typically relies on the asymptotic normality of the estimates, with the so-called sandwich formula for the variance (e.g., Kalbfleisch and Lawless 1988). Unfortunately, for our problem, deriving the sandwich formula analytically is too complicated. So in our examples we use the bootstrap method on the grouped family data. Under the bootstrap sampling, the total frequencies of the grouped data have a multinomial distribution. Since, conditional on the sum, the collection of Poisson variates has multinomial distribution, we can approximate the bootstrap samples by generating Poisson variates with means given by the observed frequencies. This means we can generate the bootstrap samples of the grouped data quite fast. We use $B = 25$ bootstrap replicates, which are sufficient since we will only use the bootstrap to compute standard errors.

Simulation study

We will address three issues by simulation: (1) robustness, (2) efficiency and (3) case definition. Previous studies

(Glidden and Liang 2002; Noh et al. 2005) indicated that the exact likelihood analysis of ascertained data is sensitive to mis-specification of the random-effect distribution. Noh et al. (2005) used a complex procedure based on hierarchical likelihood to perform a robust analysis. Here we show that robustness can also be achieved with the standard analysis if we also ascertain some proportion of the control group. Furthermore, even though the procedure in Noh et al. is robust, there is a severe loss of information from the case-only design. We show that we can retain most of the information in the full cohort by sampling all case families and a fraction of the control families. In our simulation we will compare a case-only vs case–control designs. In addition, we compare exact and pseudo-likelihood approaches for the ascertained data.

Typically a case family is defined as having at least one affected member. However, intuitively, information about variance components is captured by familial clustering, i.e., at least two affected members in the family. In our SGA problem we will also get a great computational advantage from this definition of a case family, as it leads to a substantial reduction in the number of case-family configurations. In the simulation, we will compare the efficiency of the estimates under different definitions of case family (at least one affected vs at least two affected members).

Following the example in Noh et al. (2005), we simulated a population of 100,000 families, each comprising $n_i = 5$ siblings (children only, parents not included). Additionally, considering the small family sizes in the real data, we also simulated families with $n_i = 3$ siblings. The binary outcomes are assumed Bernoulli with probability π_{ij}, which follows the logistic mixed-model

$$\log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = x'_{ij}\beta + b_i,$$

where b_i is assumed $N(0, \theta)$. We define two family-level covariates (i.e. fixed across members within a family), both generated to follow the standard uniform distribution. (Having siblings within a family share the same covariates simplify the computations, but with real data this is not necessary.) The fixed parameters are set at $\beta_0 = -0.5$, $\beta_1 = 0.15$, $\beta_2 = 0.20$. Given these assumptions, the basic likelihood contribution from the ith sibship is

$$f(y_i, x_i, \beta, \theta) = \int_{b_i} \binom{n_i}{d_i} \pi_i^{d_i}(1-\pi_i)^{n_i-d_i}\phi(b_i)db_i,$$

where $\pi_i \equiv \pi_{ij}$, and $\phi(b_i)$ is the normal density for b_i. The integral can be computed very fast using the Gaussian quadrature method.

To assess the robustness against mis-specification, we generate the random effects b_i according to each of the following distributions:

1. $b_i \sim N(0, 4.5)$,
2. $b_i \sim$ Logistic with mean 0 and variance 4.5.

The first model provides a partial check on the simulation, where the exact method should work well. The second is a heavy-tailed model, which has been shown to produce biased estimates (Noh et al. 2005). In both cases, b_i has a true variance $\theta = 4.5$. (We also tried another heavy-tailed model $b_i\sqrt{2.7}t(5)$, where $t(5)$ is the t-distribution with 5 degrees of freedom, and a skewed model $b_i\sqrt{2.25}(u_i - 2)$, where u_i follows the gamma distribution with shape parameter 2 and scale parameter 1. The results were similar and will not be shown here)

The means and standard deviations of the parameters estimates from 200 datasets are presented in Table 1. When the model is correct, i.e., the true random-effect distribution is normal, all procedures are consistent (scenarios A1 to A4). However, note the substantial increase of variance in the case-only design (A2). For estimation of θ in A2, the efficiency vs the full cohort is $(0.085/0.306)^2 = 0.08$. Even when only 5% of the control families are included A3, the efficiency is $(0.085/0.099)^2 = 0.75$. Lower efficiency is achieved for the regression parameters, but this can be improved by increasing the sampling proportion of the controls. Furthermore, the results also indicate that the pseudo-likelihood (A4) achieves similar results as the exact likelihood (A3).

If the true random effects are not normally distributed, but the model still assumes normality, then the case-only design (A6) can produce very misleading estimates. This problem was first presented in Glidden and Liang (2002), and investigated further in Noh et al. (2005). Also, the variances of the estimates are substantially larger than those from the full cohort. In contrast, analysis of the full cohort (A5) is quite resistant to the mis-specification. More importantly, by including 5% of the control families, both the exact and pseudo-likelihood analysis of the ascertained data (A7 and A8) produce very close results to those from the full cohort data. Again, the pseudo-likelihood (A8) achieves high efficiency compared to the exact likelihood (A7).

When we change the case-family definition from at least 1 to at least 2 affected members, the number of case families drops substantially from around 13,700 to 4,400, while the number of control families increases from around 86,300 to 95,600. To achieve a similar number of ascertained families as in panel A, we increase the proportion of controls to 10% (scenarios B3 and B4). Case-only analysis (B2 and B6) continue to have robustness problems. For the case–control designs, we obtain similar results for robustness and efficiency as obtained in panel A. Finally, Table 2 shows similar results for small family size $n_i = 3$.

Table 1 Means and standard deviations from 200 simulations for full and different ascertained samples, assuming normal random effects the in logistic mixed-effect model, and using $n_i = 5$ siblings per family

Sampling design	β_0	β_1	β_2	θ
True value	−5	0.15	0.20	4.5
A. Case family: ≥1 affected				
True distribution: normal				
1. Full data	−4.998 (0.033)	0.150 (0.025)	0.197 (0.024)	4.500 (0.085)
2. Case only	−5.092 (0.241)	0.152 (0.081)	0.196 (0.078)	4.625 (0.306)
3. 5% control-exact	−4.997 (0.053)	0.151 (0.043)	0.195 (0.040)	4.501 (0.099)
4. 5% control-pseudo	−4.996 (0.055)	0.152 (0.044)	0.195 (0.043)	4.498 (0.099)
True distribution: logistic				
5. Full data	−5.381 (0.036)	0.153 (0.027)	0.202 (0.025)	5.792 (0.104)
6. Case only	−9.435 (0.540)	0.002 (0.658)	−0.07 (0.806)	11.459 (0.551)
7. 5% control-exact	−5.405 (0.054)	0.137 (0.046)	0.186 (0.044)	5.851 (0.113)
8. 5% control-pseudo	−5.380 (0.056)	0.149 (0.049)	0.201 (0.048)	5.794 (0.113)
B. Case family: >1 affected				
True distribution: normal				
1. Full data	−4.997 (0.033)	0.15 (0.023)	0.200 (0.025)	4.496 (0.079)
2. Case only	−5.220 (1.123)	0.137 (0.205)	0.221 (0.191)	4.739 (1.168)
3. 10% control: exact	−5.002 (0.054)	0.149 (0.034)	0.201 (0.035)	4.513 (0.122)
4. 10% control: pseudo	−4.997 (0.085)	0.149 (0.056)	0.200 (0.054)	4.499 (0.172)
True distribution: logistic				
5. Full data	−5.378 (0.038)	0.154 (0.029)	0.203 (0.028)	5.766 (0.108)
6. Case only	−6.371 (0.565)	0.037 (0.28)	0.073 (0.292)	8.226 (0.603)
7. 10% control: exact	−5.750 (0.081)	0.152 (0.046)	0.204 (0.044)	6.812 (0.216)
8. 10% control: pseudo	−5.382 (0.099)	0.161 (0.066)	0.207 (0.062)	5.760 (0.222)

In summary, from the simulation study we learn three things that are directly relevant in our current problem:

- Inclusion of control families increase the robustness against mis-specification of the random-effect distribution.
- In this logistic mixed-model setting, the pseudo-likelihood has high efficiency vs the exact likelihood.
- We can define case families as those having at least two affected members with little loss of information/efficiency.

Optimal matching

In our experience with the real SGA data, a direct application of the suggested sampling approach for the family case–control data does not work well: it often produces estimates that are very far from the full-likelihood estimates. This is mainly because the sampled data are often too different from the full data with respect to certain features that reflect the parameters of interest. Since the full data are available, we devise a scheme to match these features in the full data with the same features in the ascertained data.

The vector of unknown parameters can be divided into two groups: regression parameters and variance components. Hence, there are two types of statistics that are natural for matching:

- Estimates from an ordinary generalized linear model (GLM) (without the random effects),
- Odd-ratios (ORs, between family members) that capture familial risk.

We have shown previously (Yip et al. 2008) that ORs describing risk in relatives are good proxy measures of the magnitude of variance components. So if the ORs from the sampled data are similar to the ORs from the full data, then we would expect the estimates of variance components from the two datasets to be of the same magnitude. Similar thinking applies to the estimation of the regression parameters. It is of course important that the estimation of the ordinary GLM and ORs can be done extremely fast even for the full data, so we perform the following scheme:

1. Sample case and control families from the full data with the desired ascertainment probabilities.

Table 2 Means and standard deviations from 200 simulations for full and different ascertained samples, assuming normal random effects the in logistic mixed-effect model

Sampling design	β_0	β_1	β_2	θ
True value	−5	0.15	0.20	4.5
A. Case family: ≥1 affected				
True distribution: normal				
1. Full data	−5.003 (0.035)	0.151 (0.026)	0.202 (0.028)	4.508 (0.092)
2. Case only	−5.074 (0.483)	0.156 (0.101)	0.201 (0.096)	4.598 (0.6)
3. 5% control-exact	−5.004 (0.051)	0.154 (0.039)	0.202 (0.041)	4.507 (0.1)
4. 5% control-pseudo	−5.004 (0.052)	0.154 (0.041)	0.202 (0.043)	4.507 (0.1)
True distribution: logistic				
5. Full data	−5.444 (0.045)	0.149 (0.029)	0.206 (0.03)	6.002 (0.133)
6. Case only	−9.356 (0.334)	0.166 (0.307)	0.223 (0.33)	11.483 (0.332)
7. 5% control-exact	−5.004 (0.06)	0.154 (0.046)	0.202 (0.047)	4.507 (0.138)
8. 5% control-pseudo	−5.004 (0.059)	0.154 (0.048)	0.202 (0.048)	4.507 (0.138)
B. Case family: ≥2 affected				
True distribution: normal				
1. Full data	−5.006 (0.038)	0.149 (0.027)	0.2 (0.025)	4.523 (0.099)
2. Case only	−5.221 (0.534)	0.169 (0.325)	0.179 (0.351)	4.733 (0.477)
3. 10% control-exact	−5.002 (0.073)	0.149 (0.045)	0.203 (0.039)	4.514 (0.159)
4. 10% control-pseudo	−5.007 (0.111)	0.154 (0.069)	0.209 (0.06)	4.51 (0.222)
True distribution: logistic				
5. Full data	−5.442 (0.044)	0.15 (0.031)	0.204 (0.029)	5.998 (0.136)
6. Case only	−3.909 (0.48)	0.034 (0.328)	0.068 (0.313)	5.333 (0.69)
7. 10% control-exact	−5.731 (0.1)	0.149 (0.048)	0.199 (0.044)	6.791 (0.278)
8. 10% control-pseudo	−5.449 (0.112)	0.156 (0.067)	0.203 (0.069)	6.011 (0.285)

This is the same as Table 1, except here we have $n_i = 3$ siblings per family

2. Obtain ordinary GLM estimates and ORs from the full data and from the sampled data, where the latter estimates account for the ascertainment.
3. Compute the criteria

$$Q_1 = \frac{1}{h}\sum_{k=1}^{h}\frac{(\hat{OR}_k^{samp} - \hat{OR}_k^{full})^2}{\mathrm{var}(\hat{OR}_k^{samp})}$$

$$Q_2 = \frac{1}{p}\sum_{k=1}^{p}\frac{(\hat{\beta}_k^{samp} - \hat{\beta}_k^{full})^2}{\mathrm{var}(\hat{\beta}^{samp})}$$

where h is the number of ORs and p is the number of covariates. Combine the criteria into $Q = Q_1 + Q_2$ from each sampled data.

4. Repeat the procedure a large number of times, and select the sampled data that minimizes Q. In our examples, the best sample was chosen from 1000 samples. Once the sample is chosen, the estimation of the mixed model is based on the weighted likelihood (6).

While the ascertainment process looks complex, the principle is quite simple, i.e. we try to ascertain 'balanced' samples, where the balance is determined by the ORs and regression coefficients that are observed in the full data. In general, the process belongs to a stratified or two-stage sampling method. Had the data been much simpler, e.g. consisting only of families of size two and the condition involves only a single OR, then the ascertainment process becomes more transparent. In this situation, we ascertain all the case families, then sample the controls such that the ratio of cases to controls is the same as in the full data, thereby preserving the observed OR in the full data.

Since we use the bootstrap method for computing the standard errors, the complex matching does not present any analytical problem. We note that there are two ways to bootstrap the data: before or after the ascertainment step. In the former, we boostrap the full data and include the optimal matching step in the bootstrap. However, if we treat the ascertained data as a stratified sample, then it should be possible also to apply the bootstrap to the ascertained data, so the optimal matching is performed only once (to generate the ascertained data). We show later (Sect. 5) that these two methods in fact produce similar results.

Application to SGA data

The purpose of our analysis of the SGA data is to extend the results in Svensson et al. (2006) by including potential

risk factors, such as birth order, maternal smoking and maternal body-mass index (BMI). We fit model (1), which includes 4 random components: maternal, fetal, couple environment and sibling environment effects. To assess the confounding between these risk factors and the genetic and environmental effects, we first fit a simple model that includes only birth order (first = 0, subsequent = 1). In the second model we include information on preeclampcia (yes = 1, no = 0), smoking (yes = 1, no = 0) and BMI (low, medium and high).

For the purpose of matching, the data are categorized by the type of sibling pairs and the number of offspring. The sib-pair types are sister–sister, brother–sister and brother–brother, but the last two can be combined since they have the same covariance structure (Pawitan et al. 2004). From each category we compute within-sib and between-sib ORs. Irrespective of the sib-pair type (sister–sister, sister–brother, brother–brother), within-sib ORs capture the maternal and couple environment effects. In the sister–sister pairs, between-sib ORs capture the maternal and sibling environment effects. In the brother–sister and brother–brother pairs, between-sib ORs capture fetal and sibling environment effects. Table 3 shows the descriptive statistics and the ORs from the full data and the optimally-matched sample. All ORs are very well-matched between the two datasets.

To illustrate that matching is indeed necessary, the boxplots in Fig. 1 show a substantial variation between the ORs from the 1,000 randomly ascertained samples. Many samples produce ORs that are far from the corresponding full-data ORs (8.51 and 1.33 from the last category in Table 3). Without any matching, this large variability will lead to larger uncertainty in the estimates obtained from a single ascertained sample.

The results of various analysis are presented in Table 4. There is a higher risk of SGA on first or preeclamptic

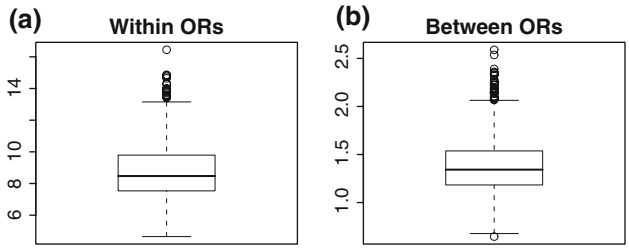

Fig. 1 Boxplots of the ORs computed from (**a**) within and (**b**) between the sibling families. The ORs in each boxplot are based on 1,000 randomly ascertained samples from the last category in Table 3

pregnancies, for smokers, or for women who have low BMI. Because of the large size of the data, all fixed regression parameters are very precisely estimated with very small standard errors. However, the estimation of the variance components, particularly the fetal component, is much less precise. With the simple model, there are only 185 observed family configurations, so there is no need for case–control sampling.

The result of the simple model is similar to Svensson et al. (2006), with a substantial fetal genetic component for SGA, accounting for a much larger contribution than the other three effects. (The result is not exactly the same, because of the different followup periods; see Sect. 2.1.) The genetic variance component parameters can be interpreted in terms of heritability (Pawitan et al. 2004), for example

$$h_m^2 = \left(\frac{0.51}{0.51 + 1.16 + 0.33 + 0.01 + 1}\right) = 16.9\%,$$

is the heritability of the maternal genetic effect. The contributions to total liability from the fetal, common family and sibling environments are 38.5, 11.0, and 0.3%, respectively. The total genetic effect (maternal + fetal)

Table 3 Counting statistics in various categories defined by sib-pair type, total number of offspring and number of offspring of the second sibling

Sib type	Offsp. total	Offsp. sib-2	No. pairs	No. case-fams	Full-data ORs		Sampled-data ORs	
					Within	Between	Within	Between
ss	2	1	32,830	35	–	1.54	–	1.55
bs, bb	2	1	96,763	88	–	1.24	–	1.24
ss	3	2	32,138	142	8.51	2.10	8.50	2.02
bs, bb	3	2	93,267	349	8.53	1.15	8.39	1.16
ss	4	1	13,114	94	7.57	1.64	7.65	1.65
bs, bb	4	1	37,112	249	8.99	1.50	8.95	1.49
ss	4	2	15,389	44	11.36	3.20	12.40	3.63
bs, bb	4	2	16,016	115	8.51	1.33	8.64	1.39
Total			326,629	1,116				

Also shown are the corresponding ORs from the full data and the optimally-matched sampled data. Within and between ORs are computed from pregnancy outcomes within and between the sib families. 'ss', 'bs' and 'bb' refer to sister–sister, brother–sister and brother–brother

Table 4 Summaries of the SGA data analysis

Variable	Full data	Full data	Sampled data	Boot2 SE	Naive SE
Regression parameters					
Constant	−1.84 (0.00)	−2.00 (0.00)	−1.98 (0.01)	(0.01)	(0.00)
Subsequent birth	−0.30 (0.00)	−0.28 (0.00)	−0.28 (0.01)	(0.00)	(0.00)
Smoking	–	0.40 (0.01)	0.41 (0.01)	(0.01)	(0.01)
Preeclampsia	–	0.88 (0.01)	0.89 (0.02)	(0.03)	(0.01)
Low vs. med BMI	–	0.22 (0.01)	0.23 (0.03)	(0.02)	(0.01)
High vs. med BMI	–	−0.08 (0.01)	−0.08 (0.01)	(0.01)	(0.01)
Variance components					
Maternal	0.51 (0.08)	0.31 (0.04)	0.29 (0.06)	(0.06)	(0.04)
Fetal	1.16 (0.44)	0.80 (0.21)	0.80 (0.16)	(0.17)	(0.20)
Couple	0.33 (0.05)	0.33 (0.04)	0.29 (0.04)	(0.03)	(0.04)
Sibling	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	(0.01)	(0.01)

The entries are the parameter estimates and their standard errors (SEs). The SEs from the full data are computed from the full likelihood. The SEs from the sampled data are computed using the bootstrap of the full data. The 'Boot2 SEs' are computed by bootstrapping the ascertained data. The 'Naive SEs' are obtained from the weighted likelihood. The value '0.00' means 'less than 0.005'

explains 55.4 (16.9 + 38.5) per cent of the liability to SGA.

When we introduce more covariates, the number of family configurations increases to 11,151, of which 802 are case-family configurations. We ascertain three-times as many control configurations as case configurations, so we achieve substantial data reduction by this sampling, while obtaining results that are very close to those of the full data. (We also tried the same number of control and case configurations, and the estimates were also quite close, except for the fetal variance.) The standard errors (SEs) of the regression estimates from the full-data bootstrap are generally larger than the SEs from the full likelihood. However, for the variance parameters the SEs are comparable. Furthermore, the SEs obtained from bootstrapping the ascertained data are comparable to the full-data bootstrap, suggesting that there is no need to bootstrap the optimal matching step. The 'naive SEs', computed from the weighted likelihood, appear too optimistic for the regression estimates, but quite comparable to the bootstrap SEs for the variance-component parameters. In practice, since the naive SEs are more readily available, we might consider using them as a first approximation, particularly for the variance-component parameters.

While still significant, in the more complex model the maternal and fetal genetic variance components drop substantially. This reflects some confounding between these components and the risk factors, which is not surprising. For example, preeclampsia is associated with both maternal and fetal genetic effects (Pawitan et al. 2004). We also expect maternal BMI to have some genetic component. The result here indicates that there are further maternal and fetal genetic effects beyond those already explained by the risk factors.

Finally, we add two more covariates: (1) maternal country of birth (Nordic = 1, other = 0), and (2) maternal age at delivery (< 26, 26–32 and > 32). The number of configurations is now 67,997, of which 1,082 are case-family configurations. Now a full-data analysis is no longer practical, particularly when numerous exploratory analysis are needed. As before, we ascertain three times as many control configurations as case configuration, and obtain the following estimates for the four variance components: 0.33(SE = 0.03), 0.80(0.20), 0.25(0.04) and 0.01(0.01). These estimates are close to those found in Table 4, which means that the two extra covariates are not confounding the genetic and environmental effects.

Conclusions

Our work on population-based family data has been motivated by questions in genetic epidemiology, particularly in estimating the relative contribution of genetic and environmental components to human diseases. Previous works in this area have been hampered by the inability to include general covariates, mainly due to computational problems in dealing with the integration of marginal likelihood. In this paper we investigated a novel approach to sampling informative families with at least two affected members, together with control families. We showed that inclusion of controls is important to preserve the robustness of the full cohort data against model mis-specification. We also showed that the pseudo-likelihood approach leads to efficient computations and the statistical properties compared well to those of the exact likelihood approach.

In the application to SGA data, the more complex model reveals more insight into the contribution of genetic factors

to this condition. For example, comparing one covariate (birth order) model with the more complex model (birth order, preeclampcia, maternal BMI, smoking) we found the total genetic contribution of liability to SGA drop from 55.4 to 45.5 per cent. This means that the genetic contribution to SGA is mostly independent of the known covariates. Similar comparison is very useful, if genotyped data are available. Then comparison between models with and without known (or candidate) risk associated single-nucleotide polymorphism (SNPs) will give us insight on how much of the total genetic effect was explained by those SNPs.

It is worth noting from the SGA analysis that, while fixed-effect regression parameters can be well estimated in this large dataset, the fetal variance component has a large standard error. This highlights the need for large population-based family data for precise estimation of genetic effects, and hence practical methods for dealing with such large datasets.

The case–control study design, commonly used in medical studies to reduce cost, collects information on cases and a subsample of controls. It is well known that a case–control study has a high efficiency compared to a full cohort study. We have a similar goal here: to sample a dataset that gives parameter estimates close to those obtained from the full cohort. Existing family-based case–control methods (e.g., Lu and Wang 2002; Moger et al. 2008) are focused on estimation of the fixed regression parameters rather than the variance components, and they usually involve only simple family structures that allow only a single variance component. For fitting the complex genetic and environmental component models, existing family-based case–control methods are still not practical enough for routine use.

While our motivation has been to reduce the computation in dealing with binary traits, it is clear that the issues and methods that we investigated here can be applied more generally to other questions, for example for quantitative-trait linkage analysis (e.g. Amos 1994; Blangero et al. 2001), where both segregation and linkage is required.

One weakness in our approach is that we can only deal with categorical covariates; continuous covariates will generate too many family configurations that the procedure becomes too slow. This is also a problem with other methods that rely on computing the likelihood for each configuration. Another weakness is typical with ascertainment methods, where, because of the subsampling, there is a potential loss of efficiency compared to the full data. However, it is worth noting that our approach is particular useful during model building stage, where speed is important but full precision less so. Once we arrive at the final stage, if feasible, we can of course use the full data for analysis.

In general, our optimal matching approach is applicable to situations where some population statistics are available. Our approach is akin to balancing considerations in two-stage sampling methodology (e.g., Reilly 1996), but the simulation approach to sample selection allows much more complex stratification.

References

Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 54:535–543

Blangero J, Williams JT, Almasy L (2001) Variance component methods for detecting complex trait loci. In: Rao DC, Province MA (eds) Genetic dissection of complex traits. Academic Press, London, pp 151–182

Breslow NE, Clayton D (1993) Approximate inference in generalized linear mixed models. J Am Stat Assoc 88:9–25

Burton PR (2003) Correcting for nonrandom ascertainment in generalized linear mixed models (GLMMs), fitted using Gibbs sampling. Genet Epidemiol 24:24–35

Burton PR, Tiller KJ, Gurrin LC, Cookson WOCM, Musk AW, Palmer LJ (1999) Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. Genet Epidemiol 17:118–140

de Andrade M, Amos CI (2000) Ascertainment issues in variance component models. Geneti Epidemiol 19:333–344

Elston RC, Sobel E (1979) Sampling considerations in the gathering and analysis of pedigree data. Am J Hum Genet 31:62–69

Epstein MP, Lin X, Boehnke M (2002) Ascertainment-adjusted parameter estimates revisited. Am J Hum Genet 70:886–895

Falconer DS (1965) The inheritance of liability to certain diseases estimated from the incidence among relatives. Ann Hum Genet 29:51–76

Genz A (1992) Numerical computation of multivariate normal probabilities. J Comput Graph Stat 1:141–149

Glidden D, Liang KY (2002) Ascertainment adjustment in complex diseases. Genet Epidemiol 23:201–208

Kalbfleisch JD, Lawless JF (1988) Likelihood analysis of multistate models for disease incidence and mortality. Stat Med 7:147–160

Lee Y, Nelder JA (1996) Hierarchical generalized linear models (with discussion). J R Stat Soc B 58:619–678

Lichtenstein P, Yip BH, Björk C, Pawitan Y, Cannon TD, Sullivan PF, Hultman CM (2009) Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. Lancet 373:234–39

Lu SE, Wang MC (2002) Cohort case–control design and analysis for clustered failure-time data. Biometrics 58:764–772

Mather K, Jinks JL (1977) Introduction to biometrical genetics. Cornell University Press, Ithaca

Moger TA, Pawitan Y, Borgan O (2008) Case-cohort methods for survival data on families from routine registers. Stat Med 27:1062–1074

Neale MC, Cardon LR (1992) Methodology for genetic studies of twins and families. Kluwer Academic, Dordrecht

Nelder JA, Mead R (1965) A simplex method for function minimization. Comput J 7:308–313

Noh M, Lee Y, Pawitan Y (2005) Robust ascertainment-adjusted parameter estimation. Genet Epidemiol 29:68–75

Noh, M, Yip B, Lee Y, Pawitan Y (2006) Multicomponent variance estimation for binary traits in family-based studies. Genet Epidemiol 30:37–47

Pawitan Y, Reilly M, Nilson E, Cnattingius S, Lichtenstein P (2004) Estimation of genetic and environmental factors for binary traits using family data. Stat Med 23:449–465

Reilly M (1996) Optimal sampling strategies for two-stage Studies. Am J Epidemiol 143:92–100

Sham PC (1998) Statistics in human genetics. Arnold, London

Svensson A, Pawitan Y, Cnattingius S, Reilly M, Lichtenstein P (2006) Familial aggregation of small-for-gestational-age births: the importance of fetal genetic effects. Am J Obstet Gynecol 194:475–9

Yip B, Björk C, Lichtenstein P, Hultman C, Pawitan Y (2008) Covariance components models for multivariate binary-traits in family data analysis. Stat Med 27:1086–1105

Zeger SL, Karim MR (1991) Generalized linear models with random effects: a Gibbs sampling approach. J Am Stat Assoc 86:79–86