Estimation of genetic effects in multiple cases family studies using penalized maximum likelihood methodology

ROULA TSONAKA*

Department of Medical Statistics and BioInformatics, Leiden University Medical Center, Post Zone S5-P, PO Box 9600, 2300 RC Leiden, The Netherlands
s.tsonaka@lumc.nl

MARIEKE C. H. DE VISser

Einthoven Laboratory for Experimental Vascular Medicine, Department of Thrombosis and Haemostasis, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands

JEANINE HOUWING-DUISTERMAAT

Department of Medical Statistics and BioInformatics, Leiden University Medical Center, Post Zone S5-P, PO Box 9600, 2300 RC Leiden, The Netherlands

SUMMARY

Family studies are often used in genetic research to explore associations between genetic markers and various phenotypes. A commonly used design oversamples families enriched with the disease under study for efficient data collection and estimation. For instance, in a multiple cases family study, families are selected based on the number of affected relatives. In such cases, valid inference for the model parameters relies on the proper modeling of both the within family correlations and the outcome-dependent sampling, also known as ascertainment. A flexible modeling approach is the ascertainment-corrected mixed-effects model, but it is known to only be asymptotically identifiable, because in small samples the available data do not provide sufficient information to estimate both the intercept and the genetic variance. To deal with this issue, we propose a penalized maximum likelihood estimation procedure which reliably estimates the model parameters in small family studies by using external population-based information.

Keywords: Ascertainment; Family-based association tests; Heritability; Outcome-dependent sampling; Penalized likelihood; Prevalence.

1. INTRODUCTION

Family-based sampling designs are often used in genetic research to identify genetic variants because they are robust to population stratification by using within family controls (Laird and others, 2000). In such designs, and when the disease under study has low-prevalence, outcome-dependent sampling schemes are additionally employed to achieve cost-effectiveness, to improve efficiency in the effect estimation

*To whom correspondence should be addressed.
and increase the power in detecting rare variants (i.e. minor allele frequency < 5%). That was also the case in the motivating study, the Genetics in Familial Thrombosis (GIFT) study, which investigates the genetic basis of venous thrombosis (VT), a multi-causal disorder affecting approximately one to three per thousand individuals per year in Western Countries (van Minkelen, 2009; Koenderman and others, 2011). All VT cases younger than 46 years of age, who have been reported in 29 Anticoagulation Clinics in the Netherlands in the period 2001–2005, were interviewed by the study investigators and information was recorded on the disease status of their siblings. Families with at least two affected siblings were recruited in the GIFT study and further genotyped to contribute to the search for novel genetic risk factors for VT.

The statistical analysis of such designs is complicated by the familial correlations and the ascertainment of the families. To handle the first issue, mixed-effects models have been proved to be an attractive modeling framework. In particular, the inclusion of random effects allows one to capture familial correlations, even in complex pedigrees, by modeling the within family relationships and allows also the estimation of parameters of interest, such as effects of genetic and environmental factors and disease heritability. To account for oversampling of disease-enriched families and avoid bias, we typically work with an adjusted likelihood that conditions on the ascertainment event. Even though maximization of the ascertainment-corrected likelihood produces asymptotically consistent results, in small samples and for rare diseases it may fail to produce reliable estimates mainly for the intercept and the random-effects variance due to lack of sufficient information (Pfeiffer and others, 2001; Neuhaus and others, 2006; Zheng and others, 2010). This in turn implies that parameters of interest defined as a function of the genetic variance, such as the disease heritability and genetic risk effects at the population level, will also be biased.

To overcome these difficulties, it has been proposed in the literature to combine information from various studies. For instance, Zheng and others (2010) and Pfeiffer and others (2008) have proposed likelihood-based and estimating equation approaches that use data from multiple sources (i.e. family and population case–control data) to produce risk effect estimates and improve the power of detecting genetic associations. Even though these approaches have been shown to work well in practice, they are limited by the fact that proper case–control data are not always available to be combined with family data. Nonetheless, in many occasions partial information for the disease of interest may be available from previous studies or population-based registries, such as the disease prevalence. In this paper, we propose to take advantage of such information, using a penalized maximum likelihood approach. In particular, the basic idea behind our proposal is to introduce penalties to the likelihood and derive parameter estimates that reflect this external information for the population from which the sample is obtained. Based on our simulation study, we show that this method leads to unbiased and efficient estimates for the parameters of interest even in small family studies.

The paper is organized as follows. In Section 2, we present the mixed-effects model framework that we have adopted for the analysis of sibship data. In Section 3, we discuss the challenges in parameter estimation under the ascertainment-corrected maximum likelihood method and present our proposed penalization approach. In Section 4, we show empirically that our method enjoys nice small sample properties and in Section 5 we exemplify it in the data from the GIFT study on 201 ascertained families.

2. Mixed models for the analysis of sibship data

Let $y_{ij}$ denote the disease status for the $j$th sibling of family $i$, and $x_{ij}$ a known $p$-dimensional vector with the genotypic information and potential additional environmental factors for $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. In the mixed-models framework sibship-specific random effects $b_i$ are introduced to model the within family dependencies, and in the case of binary responses the disease risk $\pi_{ij} = \Pr(y_{ij} = 1 \mid b_i, x_{ij})$
conditional on $b_i$ and covariate information $x_{ij}$ is modeled as (Zheng and others, 2010)

$$
\log \frac{\pi_{ij}}{1 - \pi_{ij}} = x_{ij}^T \beta + b_i, \quad b_i \sim N(0, \sigma_b^2),
$$

(2.1)

where $\beta$ is the $p \times 1$ parameter vector, $\sigma_b^2$ reflects the genetic variance, and $b_i$ and $x_{ij}$ are assumed independent. The induced log-likelihood function is then

$$
\ell(\theta) = \sum_{i=1}^n \sum_{j=1}^{n_i} \log \int_{b_i} f(y_{ij} | x_{ij}, b_i, \beta) f(b_i | \sigma_b^2) \, db_i,
$$

(2.2)

where $\theta = (\beta^T, \sigma_b^2)$, $f(y_{ij} | x_{ij}, b_i, \beta)$ is the probability density function of a binomial distribution with success probability $\pi_{ij}$, and $f(b_i | \sigma_b^2)$ is the probability density function for $b_i$, which is here taken to be the normal distribution with mean 0 and variance $\sigma_b^2$.

For the estimation of the model parameters $\theta$ using maximum likelihood, we note that the integral in the definition of the log-likelihood in (2.2) does not have in general a closed-form solution. This in practice means that the estimation of $\theta$ requires a combination of numerical integration and optimization techniques. For the maximization of the log-likelihood function with respect to $\theta$ standard algorithms can be used, such as the Expectation-Maximization (EM) algorithm (Dempster and others, 1977) or the Newton–Raphson algorithm (Lange, 1999). Regarding the numerical approximation of the integrals in (2.2), the Gaussian quadrature rule can be considered.

3. ASCERTAINMENT CORRECTION

When families are not randomly sampled from the population, but instead multiple cases families are oversampled to identify susceptibility genetic factors, the maximization of log-likelihood (2.2) leads to incorrect inference. To address this problem, the likelihood function needs to be conditioned on this ascertainment event. In particular, let $A_i$ be the indicator variable for family $i$ being included in the study; namely let $A_i = I(\sum_{j=1}^{n_i} y_{ij} \geq K)$ denote that family $i$ is included in the study only if there are at least $K$ affected siblings. In the GIFT study, $K = 2$, but the likelihood can be easily generalized for any $K \geq 1$.

Using Bayes’ theorem and, under the assumption that $\Pr(A_i = 1 | y_i, X_i) = \Pr(A_i = 1 | y_i)$, we obtain

$$
L^A(\theta) = \prod_{i=1}^n f(y_i | X_i, A_i = 1, \theta) = \prod_{i=1}^n \frac{\Pr(A_i = 1 | y_i, \theta) \, f(y_i | X_i, \theta)}{\Pr(A_i = 1 | X_i, \theta)}
$$

$$
= \prod_{i=1}^n \frac{f(y_i | X_i, \theta)}{1 - \sum_{s \in S_i} \Pr(A_i = 0 | X_i, \theta)},
$$

(3.1)

where $\Pr(A_i = 1 | y_i, \theta) = 1$, $S_i$ is the set of all possible response patterns with $\sum_{j=1}^{n_i} y_{ij} < K$ for family $i$. From (3.1), it is evident that the evaluation of the ascertainment-corrected likelihood is not more complex than (2.2) because it only requires the evaluation of family-specific densities for the observed response pattern $y_i$, and all possible response patterns in $S_i$. However, (3.1) becomes computationally more demanding with increasing family size because then the number of possible patterns increases, as well.

Similar to the random sampling case, the estimation of the model parameters $\theta$ is obtained by maximizing the ascertainment-corrected log-likelihood (3.1) using the EM or a Newton-type algorithm. For our purposes, we have chosen to work with a quasi-Newton algorithm that only requires the first-order derivatives of the log-likelihood. The expressions for these derivatives are given in Section A of the supplementary material available at Biostatistics online.
3.1 Challenges in ascertainment-corrected likelihood

As mentioned in Section 1, for rare diseases the ascertainment-corrected model (2.1) is only asymptotically identifiable. In particular, for rare diseases and for small to moderate sample sizes, there is insufficient information in the data to reliably estimate both the intercept and the random-effects variance (Pfeiffer and others, 2001; Neuhaus and others, 2006; Zheng and others, 2010). For instance, in the GIFT study, where a family is selected if at least two siblings are affected, identification of these parameters is driven only by the remaining family members and thus large family and sample sizes are required to reliably estimate the model parameters. This problem is typically manifested in practice in difficulties in convergence, due to flat likelihood surfaces, and biased parameter estimates for small samples. The bias, as we show empirically later in Section 4, is mainly observed for the baseline risk parameter and the random-effects variance. This bias carries over to related quantities of interest, such as the disease heritability and population-averaged parameters, that entail the random-effects variance. Specifically, in the family setting that we consider here, the disease heritability $H^2$, which measures the percentage of variation in the phenotype that has a genetic origin, is obtained as (Houwing-Duistermaat and others, 2000)

$$H^2 = \frac{2\sigma_h^2}{2\sigma_h^2 + \pi^2/3},$$  \hspace{1cm} (3.2)

where $\sigma_h^2$ reflects the genetic variance in the mixed-effects logistic regression (2.1) and $\pi^2/3$ is the variance of the standard logistic distribution, with $\pi$ the known mathematical constant. The percentage of genetic variance attributed to a specific genetic factor is then derived as

$$h^2 = \frac{\sigma_{\beta_1=0}^2 - \sigma_h^2}{\sigma_{\beta_1=0}^2},$$  \hspace{1cm} (3.3)

where $\sigma_{\beta_1=0}^2$ is the random-effects variance in a mixed-effects logistic regression without including the fixed effect of the genetic risk factor in the linear predictor, and $\sigma_h^2$ is the random-effects variance from a mixed model that incorporates the effect of this factor. It is obvious from (3.2) and (3.3) that since the heritability measures are defined as a function of the random-effects variance, potential bias in estimating this variance component carries over to $H^2$ and $h^2$.

Despite the bias in the baseline risk and genetic variance, as we show in our simulation study in Section 4, the estimation of genetic or environmental effects is not affected. However, note that the parameters $\beta$ in model (2.1) have an interpretation conditional on the sibship-specific random effects, which is not always desirable. This in practice means that we cannot generalize the obtained inference to the population from which the sibships are sampled. This is mainly because under a non-linear link function, such as the logit link, $\int \logit^{-1}(x_i^T \beta + b_i) dF(b_i) \neq \logit^{-1}(x_i^T \beta)$, where $\logit(x) = \log(x/(1-x))$, $\logit^{-1}(x) = \exp(x)/(1 + \exp(x))$, and $F(.)$ denotes the cumulative distribution function for $b_i$.

Nonetheless, in the special case of the sibships that we consider here, it is possible from model (2.1) to derive parameters with the desired population-averaged interpretation. In particular, under the random-intercepts setting Zeger and others (1988) have shown that $\logit[ E(y_{ij} | x_{ij}) ] = \logit[\pi(\theta, x_{ij})] \approx (c^2\sigma_h^2 + 1)^{-1/2}x_{ij}^T \beta$, with $c^2 = 0.346$. This means that, from $\beta$, we can obtain the parameters $\beta^M$ that have a marginal interpretation using

$$\beta^M \approx (c^2\sigma_h^2 + 1)^{-1/2} \beta,$$  \hspace{1cm} (3.4)

where we observe that the random-effects variance $\sigma_h^2$ is also involved. This result shows that even though a consistent estimate for $\beta$ (expect from the intercept) can be obtained, potential bias in the estimation of $\sigma_h^2$ will inevitably affect $\beta^M$. 

---

**References:**

Pfeiffer, Neuhaus, Zheng, Houwing-Duistermaat.
3.2 Penalized likelihood estimation

To overcome the difficulties in the analysis of ascertained families, we propose to incorporate in the estimation procedure external information, namely the disease prevalence, which is often known for most diseases by population-based registries. Let $\pi_0$ denote the value of the disease prevalence; then the model parameters can be estimated using the penalized log-likelihood

$$
\ell^{AP}(\theta) = \ell^{A}(\theta) - \lambda \| \logit(\pi(\theta)) - \logit(\pi_0) \|^2,
$$

where $\ell^{A}(\theta)$ is the logarithm of (3.1), $\lambda > 0$ is a penalty parameter and $\pi(\theta)$ is the model-based population prevalence calculated as

$$
\pi(\theta) = \Pr(y_{ij} = 1) = \int \int \Pr(y_{ij} = 1 \mid b_i, x_{ij}) f(x_{ij}) f(b_i) \, dx_{ij} \, db_i.
$$

In practice, for continuous variables $x_{ij}$ and $b_i$, the integrals in (3.6) can be approximated numerically using Gaussian quadrature methods.

In the special case of the sibships that we consider here, the integration with respect to the random effects distribution can be easily approximated using the marginal formula of Zeger and others (1988), as explained in Section 3.1. Regarding the integral over the space of $x_{ij}$, here we will assume that the disease risk is dependent only on a single genetic factor of interest in addition to the family-specific random effects $b_i$, i.e. $x_{ij} = [1, \text{snp}_{ij}]$ where $\text{snp}_{ij} \in \{0, 1, 2\}$ denotes the number of minor alleles of each sibling under an additive genetic model. In this case, (2.1) simplifies to

$$
\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_0 + \beta_1 \text{snp}_{ij} + b_i,
$$

where $\beta_0$ denotes the baseline log odds and $\beta_1$ measures the change in the log odds for a unit increase in $\text{snp}_{ij}$ conditional on the random effects $b_i$, i.e. within a particular family $i$. Assuming that the genotype frequencies $p_0, p_1, p_2$ are known or estimated by the data, with $\sum_{k \in \{0, 1, 2\}} p_k = 1$, we obtain the prevalence as $\pi(\theta) = \sum_{k \in \{0, 1, 2\}} p_k \pi(\theta, x_k)$, where $\pi(\theta, x_k) = \logit^{-1}(\beta_0^M + \beta_1^M k)$, with $\beta_0^M$ and $\beta_1^M$ given by (3.4). Clearly, when more covariates need to be incorporated in the linear predictor, the evaluation of the integral with respect to their joint distribution will inevitably become more intensive.

For the estimation of the model parameters $\theta$, we need to maximize the penalized ascertainment-corrected log-likelihood (3.5) for fixed $\lambda$. Contrary to the various penalized-likelihood techniques proposed in the literature, our goal here is not to choose a $\lambda$ value that optimizes the model’s predictive ability. Instead the value of $\lambda$ should reflect the information we have available about the prevalence in the population. In that respect, our penalized likelihood shares similarities with the Bayesian framework. In fact, the quadratic penalty used in (3.5) corresponds to putting a prior on the disease prevalence, which is a non-linear function of the model parameter $\theta$. Specifically (3.5) corresponds to assuming as prior $\logit(\pi(\theta)) \sim N(\logit(\pi_0), (2\lambda)^{-1})$. Depending on the precision of the information that is available for the disease prevalence, the value for $\lambda$ should be chosen accordingly. We distinguish two cases; first when $\pi_0$ is known exactly (i.e. from population registries), and second when $\pi_0$ is known to lie in an interval (e.g. from a previous study). In the first case, we should choose a large value for $\lambda$ in order to ensure that $\pi_0$ takes this value. This is in fact equivalent to a constrained optimization problem where the averaged disease risk (3.6) should equal the pre-specified prevalence value. Susko and others (1998) have argued that the use of a quadratic penalty results in a $\pi(\hat{\theta})$ that satisfies the constraint for fixed $\lambda$ without the need to update it in each iteration. For a sufficiently large $\lambda$, the constraint is approximately satisfied, but numerical problems can occur if $\lambda$ is chosen too large. In our experience, a $\lambda$ value of magnitude $10^{-1}|\ell(\hat{\theta})_0|$, where $\hat{\theta}_0$ denotes starting values derived from a standard mixed model (without the ascertainment correction), leads to a
\( \pi(\hat{\theta}) \) that satisfies the constraint without numerical difficulties. In the case where the prevalence is known to lie in an interval, e.g. \([a, b]\), a smaller \(\lambda\) value needs to be chosen. For this example, and from a Bayesian perspective this corresponds to a prior \(\logit(\pi(\theta)) \sim N(\mu, \sigma^2)\) with \(\mu = (\logit(a) + \logit(b))/2\) taken at the mid-point of the interval on the logit scale, and \(\sigma = (1/2\lambda)^{1/2} = (\logit(b) - \logit(a))/6\). Namely, to ensure that the estimated prevalence lies within \([a, b]\), \(\lambda\) should be chosen such that the width of this interval on the logit scale equals six times the standard deviation of this distribution. This choice is motivated by the properties of the normal distribution, namely by the fact that 99.73% of the values lie within three standard deviations of the mean. Moreover, based on the similarities of the penalized method with the Bayesian framework, we should note that using heuristic arguments from asymptotic Bayesian theory (Cox and Hinkley, 1974, pp. 399–400), asymptotically the behavior of the estimators is governed by the likelihood contribution and not by the prior. This means that asymptotically the quadratic penalty will not contribute anymore to the estimation of the model parameters.

Finally, similar to the optimization of the unpenalized log-likelihood \(\ell^A(\theta)\), a combination of numerical integration and optimization methods is required. The score vectors of (3.5) with respect to \(\theta\) are given in Section A of the supplementary material available at Biostatistics online.

4. Simulation study

In this section, we evaluate the performance of our proposed method under different simulation scenarios for disease prevalence and heritability. Motivated by the GIFT study briefly introduced in Section 1 and analyzed in Section 5, we simulate multiple cases families with sibships using the mixed-effects logistic regression (3.7) with \(b_i \sim N(0, \sigma_{b_i}^2)\) and \(j = 1, \ldots, 5\) for all sibships \(i\). For different choices for \(\beta_0\) and \(\sigma_b^2\), we set up nine simulation scenarios that correspond to \(\pi_0 = 1\%, 5\%,\) and \(10\%\) prevalence, while the heritability is taken to equal \(25\%, 50\%,\) and \(60\%\). The choices for these parameter values were motivated by the GIFT study. In particular, for VT the prevalence for people younger than 46 years of age is known to equal 1.2% (Briët and others, 1994) and the heritability has been estimated between 50% and 60%. The assumed values for \(\beta_0\) and \(\sigma_b^2\) are given in Tables 1–3 for each scenario. For the SNP effect, we assumed \(\beta_1 = 0.5\) in all scenarios and the minor allele frequency was set at 30%. For each scenario, 1000 datasets are simulated and for each dataset \(n = 100\) and \(n = 500\) families with at least two affected siblings are selected for further analysis. We estimated the model parameters by maximizing both the unpenalized ascertainment-corrected likelihood (3.1) and the penalized likelihood (3.5). The genotypic frequencies required for the evaluation of the penalty term in (3.5) are estimated from the collected samples. In our simulations, we have also evaluated the impact of the chosen penalty parameter, the effect of prevalence uncertainty and misspecification and we illustrated the performance of the method when extra covariates are added in the model. All results of this simulation study are presented in Section B of the supplementary material available at Biostatistics online. A part of these results is given in Tables 1–3, where to enhance comparisons between the penalized and unpenalized analyses, we present the mean of the estimated parameters over the 1000 datasets and in brackets the root mean square error (RMSE) and 95% coverage probabilities (CPr).

It is evident from Tables 1–3 that the unpenalized likelihood (3.1) suffers from severe bias in both the intercept and the random-effects variance, especially as heritability increases. However, as we have expected when we take advantage of the disease prevalence information, the RMSE is considerably improved. Moreover, as discussed in Section 3.1, bias in the estimation of the intercept and the random-effects variance carries over to other quantities of interest, i.e. \(\beta^M\) and \(H^2\) and therefore RMSE is higher for the unpenalized approach. It is also the 95% CPr that are affected since the unpenalized method fails to preserve them at a nominal level for most parameters. In addition, when we used the unpenalized likelihood for the three prevalence value levels (i.e. 1%, 5%, and 10%), on average 14.8%, 22.2%, and 1.3% of the simulated datasets failed to converge, whereas no convergence issues were observed for the penalized likelihood.
Table 1. Simulation study results: Comparisons between the unpenalized ascertainment-corrected likelihood (3.1) with the penalized one (3.5) for disease prevalence 1% and different choices for disease heritability (i.e. \( H^2 = 0.23, 0.48, 0.58 \))

<table>
<thead>
<tr>
<th>Pars</th>
<th>True</th>
<th>Est (RMSE) (CPr)</th>
<th>True</th>
<th>Est (RMSE) (CPr)</th>
<th>True</th>
<th>Est (RMSE) (CPr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_b )</td>
<td>0.71</td>
<td>0.793 (0.149) (0.951)</td>
<td>1.23</td>
<td>0.926 (0.442) (0.845)</td>
<td>1.50</td>
<td>1.076 (0.574) (0.740)</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>-5.32</td>
<td>-6.132 (1.957) (0.964)</td>
<td>-6.00</td>
<td>-5.005 (1.664) (0.671)</td>
<td>-6.40</td>
<td>-4.728 (2.227) (0.621)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.50</td>
<td>0.491 (0.189) (0.967)</td>
<td>0.50</td>
<td>0.500 (0.177) (0.660)</td>
<td>0.50</td>
<td>0.500 (0.198) (0.606)</td>
</tr>
<tr>
<td>( \beta_1^M )</td>
<td>0.46</td>
<td>0.444 (0.172) (0.952)</td>
<td>0.41</td>
<td>0.437 (0.162) (0.960)</td>
<td>0.38</td>
<td>0.420 (0.174) (0.951)</td>
</tr>
<tr>
<td>( H^2 )</td>
<td>0.23</td>
<td>0.275 (0.070) (0.951)</td>
<td>0.48</td>
<td>0.331 (0.207) (0.845)</td>
<td>0.58</td>
<td>0.395 (0.251) (0.740)</td>
</tr>
</tbody>
</table>

The mean of the estimated parameters with the ‘RMSE’ and 95% ‘CPr’ in brackets, is presented for the square root of the random-effects variance \( (\sigma_b) \), the baseline log odds \( (\beta_0) \), the family-specific SNP effect \( (\beta_1) \), the population-averaged SNP effect \( (\beta_1^M) \), and the heritability \( (H^2) \).

Table 2. Simulation study results: Comparisons between the unpenalized ascertainment-corrected likelihood (3.1) with the penalized one (3.5) for disease prevalence 5% and different choices for disease heritability (i.e. \( H^2 = 0.23, 0.48, 0.58 \))

<table>
<thead>
<tr>
<th>Pars</th>
<th>True</th>
<th>Est (RMSE) (CPr)</th>
<th>True</th>
<th>Est (RMSE) (CPr)</th>
<th>True</th>
<th>Est (RMSE) (CPr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma_b )</td>
<td>0.71</td>
<td>0.829 (0.326) (0.836)</td>
<td>1.23</td>
<td>1.091 (0.419) (0.813)</td>
<td>1.50</td>
<td>1.242 (0.468) (0.902)</td>
</tr>
<tr>
<td>( \beta_0 )</td>
<td>-3.53</td>
<td>-4.117 (1.369) (0.882)</td>
<td>-3.96</td>
<td>-3.734 (1.347) (0.756)</td>
<td>-4.25</td>
<td>-3.525 (1.394) (0.817)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.50</td>
<td>0.508 (0.188) (0.977)</td>
<td>0.50</td>
<td>0.497 (0.193) (0.749)</td>
<td>0.50</td>
<td>0.496 (0.186) (0.798)</td>
</tr>
<tr>
<td>( \beta_1^M )</td>
<td>0.46</td>
<td>0.454 (0.171) (0.946)</td>
<td>0.41</td>
<td>0.416 (0.163) (0.951)</td>
<td>0.38</td>
<td>0.399 (0.154) (0.955)</td>
</tr>
<tr>
<td>( H^2 )</td>
<td>0.23</td>
<td>0.286 (0.146) (0.836)</td>
<td>0.48</td>
<td>0.401 (0.189) (0.813)</td>
<td>0.58</td>
<td>0.464 (0.199) (0.902)</td>
</tr>
</tbody>
</table>

The mean of the estimated parameters with the ‘RMSE’ and 95% ‘CPr’ in brackets, is presented for the square root of the random-effects variance \( (\sigma_b) \), the baseline log odds \( (\beta_0) \), the family-specific SNP effect \( (\beta_1) \), the population-averaged SNP effect \( (\beta_1^M) \), and the heritability \( (H^2) \).

(3.5). When \( n = 500 \), the RMSE and 95% CPr are considerably improved compared with the \( n = 100 \) case for both the unpenalized and penalized methods (Tables S10–S12 of the supplementary material available at Biostatistics online). In addition, when the sample size increases, no dataset fails to converge (Table S14 of the supplementary material available at Biostatistics online).
our goal is first to estimate the disease heritability and then study the association of FVL and O with disease risk. To this end, we use the mixed-effects logistic regression (3.7), where $\text{snpi}_j$ here denotes the genotypes for FVL or the blood group O and $h_j \sim N(0, \sigma^2_h)$. Here we should note that genetic information has been recorded only for siblings who were alive at the data collection. Therefore, this means that inference using the ascertainment-corrected likelihood (3.1) is valid under the assumption of the missing at random (MAR) mechanism. In fact, such an assumption for the GIFT study is realistic because a young group is considered, and thus there is a very low probability to dying of VT. However, in studies where a MAR assumption does not hold, the analysis should be adjusted for it using joint models for the measurement and missingness process (Molenberghs and Kenward, 2007).

### Table 3. Simulation study results: Comparisons between the unpenalized ascertainment-corrected likelihood (3.1) with the penalized one (3.5) for disease prevalence 10% and different choices for disease heritability (i.e. $H^2 = 0.23$, 0.48, 0.58)

<table>
<thead>
<tr>
<th>Pars</th>
<th>True</th>
<th>Est (RMSE) (CPr)</th>
<th>True</th>
<th>Est (RMSE) (CPr)</th>
<th>True</th>
<th>Est (RMSE) (CPr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unpenalized method—prevalence 10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sigma_b$</td>
<td>0.71</td>
<td>0.861 (0.372) (0.865)</td>
<td>1.23</td>
<td>1.133 (0.416) (0.824)</td>
<td>1.50</td>
<td>1.309 (0.445) (0.934)</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>-2.71</td>
<td>-3.306 (1.327) (0.930)</td>
<td>-3.03</td>
<td>-2.961 (1.162) (0.793)</td>
<td>-3.25</td>
<td>-2.860 (1.134) (0.836)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.50</td>
<td>0.516 (0.189) (0.964)</td>
<td>0.50</td>
<td>0.499 (0.188) (0.787)</td>
<td>0.50</td>
<td>0.498 (0.187) (0.816)</td>
</tr>
<tr>
<td>$\beta^M_1$</td>
<td>0.46</td>
<td>0.457 (0.169) (0.946)</td>
<td>0.41</td>
<td>0.413 (0.158) (0.950)</td>
<td>0.38</td>
<td>0.392 (0.149) (0.954)</td>
</tr>
<tr>
<td>$H^2$</td>
<td>0.23</td>
<td>0.300 (0.166) (0.865)</td>
<td>0.48</td>
<td>0.418 (0.183) (0.824)</td>
<td>0.58</td>
<td>0.488 (0.186) (0.934)</td>
</tr>
</tbody>
</table>

| Penalized method—prevalence 10% | | | | | | |
| $\sigma_b$ | 0.71 | 0.724 (0.121) (0.925) | 1.23 | 1.227 (0.153) (0.954) | 1.50 | 1.496 (0.165) (0.942) |
| $\beta_0$ | -2.71 | -2.750 (0.166) (0.957) | -3.03 | -3.044 (0.184) (0.963) | -3.25 | -3.253 (0.197) (0.946) |
| $\beta_1$ | 0.50 | 0.508 (0.186) (0.961) | 0.50 | 0.502 (0.188) (0.954) | 0.50 | 0.504 (0.187) (0.944) |
| $\beta^M_1$ | 0.46 | 0.467 (0.171) (0.946) | 0.41 | 0.407 (0.153) (0.944) | 0.38 | 0.379 (0.141) (0.946) |
| $H^2$ | 0.23 | 0.242 (0.060) (0.925) | 0.48 | 0.475 (0.062) (0.954) | 0.58 | 0.572 (0.055) (0.942) |

The mean of the estimated parameters with the ‘RMSE’ and 95% ‘CPr’ in brackets, is presented for the square root of the random-effects variance ($\sigma_b$), the baseline log odds ($\beta_0$), the family-specific SNP effect ($\beta_1$), the population-averaged SNP effect ($\beta^M_1$), and the heritability ($H^2$).
We found that, using the unpenalized ascertainment-corrected likelihood (3.1), formal decision. a dominant model may seem more appropriate, but there is certainly not enough information to make a account. Regarding FVL, we have used a codominant genetic model, but based on the estimated odds ratios observe that the marginal odds ratios are reduced when knowledge on the disease prevalence is taken into leads to a considerably narrower CI for both the conditional and marginal odds ratios. In addition, we RMSE for the genetic effect remains at the same level in both methods. However, the penalized analysis both methods. This agreement is in fact expected because, as we have seen in our simulation study, the this analysis, the heritability was estimated as 0.588 with 95% CI (0.536, 0.638).

For the estimation of disease heritability, the genetic factor is excluded from the linear predictor in (3.7). We found that, using the unpenalized ascertainment-corrected likelihood (3.1), $H^2$ is 0.209 with 95% confidence interval (CI) (0.086, 0.426), and the estimated disease prevalence is 9.6% (95% CI: [4.9%; 17.9%]), which is considerably higher than 1.2%. This overestimation in prevalence suggests that the heritability is potentially underestimated as well. To further investigate this, we analyzed the GIFT data using the penalized ascertainment-corrected likelihood (3.5) which takes advantage of the prevalence of VT. Based on this analysis, the heritability was estimated as 0.588 with 95% CI (0.536, 0.638).

We proceed to the estimation of the effect of FVL variant and blood group O using both the unpenalized and penalized likelihood. The parameter estimates with corresponding 95% asymptotic normal CIs are given in Table 4. From our analysis, the association of both FVL and O with VT is corroborated using both methods. This agreement is in fact expected because, as we have seen in our simulation study, the RMSE for the genetic effect remains at the same level in both methods. However, the penalized analysis leads to a considerably narrower CI for both the conditional and marginal odds ratios. In addition, we observe that the marginal odds ratios are reduced when knowledge on the disease prevalence is taken into account. Regarding FVL, we have used a codominant genetic model, but based on the estimated odds ratios a dominant model may seem more appropriate, but there is certainly not enough information to make a formal decision.

When studying the contribution of both FVL and O to the genetic variance, it is estimated at 48.6% without penalization, whereas with penalization it drops to 1.45%. Thus, ignoring knowledge on disease prevalence considerably overestimates the percentage of genetic variance attributed to specific risk factors. In addition, such a difference between the unpenalized and penalized method has been also seen in our

| Table 4. GIFT study: Parameter estimates and 95% CIs for FVL and O, using the unpenalized method, the penalized method that assumes that prevalence is known at 1.2%, and the penalized method that assumes that prevalence lies in [0.8%, 1.6%] |
|---|---|---|---|
| | Unpenalized | Penalized | Penalized-uncertainty |
| | Estimate | CI | Estimate | CI | Estimate | CI |
| FVL | | | | | | |
| Intercept | -4.063 | (-7.461, -0.666) | -6.548 | (-9.262, -6.169) | -5.968 | (-8.684, -5.072) |
| FVL-1 | 1.663 | (1.051, 2.275) | 1.578 | (1.049, 2.108) | 1.683 | (1.111, 2.254) |
| FVL-2 | 1.835 | (-0.191, 3.861) | 1.704 | (-0.183, 3.591) | 1.826 | (-0.152, 3.804) |
| exp(FVL-1) | 5.276 | (2.862, 9.728) | 4.847 | (2.856, 8.288) | 5.381 | (3.038, 9.529) |
| exp(FVL-2) | 6.264 | (0.826, 47.514) | 5.497 | (0.833, 36.273) | 6.211 | (0.859, 44.887) |
| exp(FVL-1)^† | 4.071 | (2.457, 6.747) | 3.237 | (2.173, 4.822) | 3.532 | (2.297, 5.431) |
| exp(FVL-2)^† | 4.706 | (0.855, 25.894) | 3.555 | (0.872, 14.499) | 3.933 | (0.892, 17.332) |
| σ_0 | 1.080 | (0.436, 2.675) | 1.526 | (1.400, 1.663) | 1.500 | (1.303, 1.727) |
| O | | | | | | |
| Intercept | -2.333 | (-4.444, -0.221) | -5.062 | (-8.977, -2.228) | -4.789 | (-6.526, -3.051) |
| O | -0.774 | (-1.280, -0.288) | -0.868 | (-1.355, -0.381) | -0.855 | (-1.346, -0.364) |
| exp(O) | 0.461 | (0.284, 0.749) | 0.419 | (0.258, 0.683) | 0.425 | (0.260, 0.695) |
| exp(O)^M | 0.489 | (0.314, 0.762) | 0.518 | (0.355, 0.757) | 0.519 | (0.351, 0.766) |
| σ | 0.709 | (0.201, 2.495) | 1.471 | (1.233, 1.753) | 1.420 | (1.054, 1.913) |

† The 95% CI for the marginal odds ratios have been computed using the Delta method.

For the estimation of disease heritability, the genetic factor is excluded from the linear predictor in (3.7). We found that, using the unpenalized ascertainment-corrected likelihood (3.1), $H^2$ is 0.209 with 95% confidence interval (CI) (0.086, 0.426), and the estimated disease prevalence is 9.6% (95% CI: [4.9%; 17.9%]), which is considerably higher than 1.2%. This overestimation in prevalence suggests that the heritability is potentially underestimated as well. To further investigate this, we analyzed the GIFT data using the penalized ascertainment-corrected likelihood (3.5) which takes advantage of the prevalence of VT. Based on this analysis, the heritability was estimated as 0.588 with 95% CI (0.536, 0.638).

We proceed to the estimation of the effect of FVL variant and blood group O using both the unpenalized and penalized likelihood. The parameter estimates with corresponding 95% asymptotic normal CIs are given in Table 4. From our analysis, the association of both FVL and O with VT is corroborated using both methods. This agreement is in fact expected because, as we have seen in our simulation study, the RMSE for the genetic effect remains at the same level in both methods. However, the penalized analysis leads to a considerably narrower CI for both the conditional and marginal odds ratios. In addition, we observe that the marginal odds ratios are reduced when knowledge on the disease prevalence is taken into account. Regarding FVL, we have used a codominant genetic model, but based on the estimated odds ratios a dominant model may seem more appropriate, but there is certainly not enough information to make a formal decision.

When studying the contribution of both FVL and O to the genetic variance, it is estimated at 48.6% without penalization, whereas with penalization it drops to 1.45%. Thus, ignoring knowledge on disease prevalence considerably overestimates the percentage of genetic variance attributed to specific risk factors. In addition, such a difference between the unpenalized and penalized method has been also seen in our
simulation study and is attributed to the unstable estimation of $\sigma_b$ (see Section B in the supplementary material available at *Biostatistics* online).

Finally, as discussed in Section 3.2, when the prevalence is not known exactly, we can incorporate into our analysis this uncertainty by making an appropriate choice for the penalty parameter $\lambda$. Specifically, for the GIFT study even though the probands were younger than 46 years of age, their siblings are older than 46 years of age in some cases (histogram of their ages presented in Section C of the supplementary material available at *Biostatistics* online). Therefore, it is reasonable to assume that in GIFT the prevalence is only known with some uncertainty and thus, here, for the sake of illustration, we will assume it ranges from 0.8% to 1.6%. According to Section 3.2, accounting for this uncertainty corresponds to maximizing the penalized likelihood using $\lambda = 36.6$. The results are shown in Table 4. Thus, we observe that, as expected, uncertainty about the disease prevalence leads to estimates that get closer to the results of the unpenalized likelihood, which ignores any prior knowledge on prevalence.

To sum up, for all the parameters of interest we have observed considerable differences between the penalized and unpenalized analyses. Taking into account the results from the simulation study, we should put more faith on the penalized analysis that takes advantage of the prior information and ensures that the derived results obey the prevalence specification that we expect in the population.

6. Discussion

We have proposed a penalized maximum likelihood estimation approach to improve the reliability of the ascertainment-corrected likelihood in studies with a few families. This method is simple to implement and, based on the results from our simulation study and analysis of GIFT data, we have shown that it can considerably improve the precision of our estimates and produces unbiased estimates of the genetic variance, even for small sample and family sizes. In addition, reliable estimates for the variance parameters lead to unbiased estimates of odds ratios at the population level, which is often of primary interest. Thus, using our proposed method, small family studies can contribute to statistically significant findings in a meta-analysis that involves various family-based or case–control studies. However, we should note that as we show empirically in Section B.3 of the supplementary material available at *Biostatistics* online, mis-specification of the disease prevalence can lead to severe bias, but nowadays information on the prevalence is available for many diseases, making our approach generally applicable.

A key component of our method is the penalty term which is derived as the expectation of the model-based prevalence with respect to the joint distribution of the involved covariates. In our analyses so far, we have made parametric assumptions for the distribution of the covariates. Nevertheless, in cases where making a parametric assumption is not appropriate, a non-parametric density estimation method, e.g. a kernel density estimator, can be used. In addition, for highly selected families, and especially when selection is dependent on these covariates, we can make a parametric assumption for the density of the covariates while correcting for truncation.

In the description of the proposed method, we have considered the prospective ascertainment-adjusted likelihood, because in the GIFT study it is possible to explicitly model the ascertainment event (i.e. at least two affected siblings). However, for more complex sampling schemes where the ascertainment is more difficult to be modeled, the retrospective likelihood is an attractive alternative. Examples include designs that recruit affected sibling pairs and unrelated controls, or sample sibships with affected parents, etc. Our proposal can be easily applied in this case as well by introducing the same penalty term in the corresponding likelihood function. The same also holds for the joint likelihood approach (Kraft and Thomas, 2000) to model the ascertainment process. Finally, another area of application of our method is when data from multiple sources (e.g. family studies with case–control data) are combined to increase the sample size and thereby produce reliable genetic effect estimates (Zheng and others, 2010; Balliu and others, 2012).
Introducing in this case external information for the disease can lead to even more efficient estimates and increase the power of detecting genetic effects.

**SUPPLEMENTARY MATERIAL**


**ACKNOWLEDGMENTS**

Conflict of Interest: None declared.

**FUNDING**

This work was supported by the Netherlands Organisation for Scientific Research [917-66-344 and 912-02-036] and the Netherlands Heart Foundation [2005T055].

**REFERENCES**


[Received December 14, 2011; revised August 2, 2012; accepted for publication August 18, 2012]