Statistical methods for the analysis of complex omics data

Tissier, R.

2

Secondary Phenotype Analysis in Ascertained Family Designs: Application to the Leiden Longevity Study

Abstract

The case-control design is often used to test associations between the case-control status and genetic variants. In addition to this primary phenotype a number of additional traits, known as secondary phenotypes, are routinely recorded and typically associations between genetic factors and these secondary traits are studied too. Analysing secondary phenotypes in case-control studies may lead to biased genetic effect estimates, especially when the marker tested is associated with the primary phenotype and when the primary and secondary phenotypes tested are correlated. Several methods have been proposed in the literature to overcome the problem but they are limited to case-control studies and not directly applicable to more complex designs, such as the multiple-cases family studies. A proper secondary phenotype analysis, in this case, is complicated by the within families correlations on top of the biased sampling design. We propose a novel approach to

This chapter has been published as: Renaud Tissier, Roula Tsonaka, Simon P. Mooijart, P. Eline Slagboom, Jeanine J. Houwing-Duistermaat (2017). Secondary Phenotype Analysis in Ascertained Family Designs: Application to the Leiden Longevity Study. *Statistics in Medicine* 36(14), 2288-2301.

accommodate the ascertainment process while explicitly modelling the familial relationships. Our approach pairs existing methods for mixed-effects models with the retrospective likelihood framework and uses a multivariate probit model to capture the association between the mixed type primary and secondary phenotypes. To examine the efficiency and bias of the estimates we performed simulations under several scenarios for the association between the primary phenotype, secondary phenotype, and genetic markers. We will illustrate the method by analysing the association between triglyceride levels and glucose (secondary phenotypes) and genetic markers from the Leiden Longevity study, a multiple-cases family study that investigates longevity.

2.1 Introduction

In order to understand biological mechanisms underlying disease and health, epidemiological studies measure genetic markers, classical variables, and novel omics datasets and model the relationship between these variables and the phenotype of interest. Here we consider outcome dependent sampling designs with binary outcome variables. In addition to studying these binary (primary) phenotypes, the classical or omics variables are typically also analysed as outcome variables (secondary phenotypes). For example modelling of associations between these traits and genetic factors, such as single-nucleotide polymorphisms (SNPs) or polygenic risk scores (sumscores based on SNPs)(Dubdbridge, 2003). However, an important complication which is often ignored is that a proper analysis of the secondary traits should correct for the sampling mechanism on the primary phenotype (Figure 2.1). Note that we assume that the secondary phenotype has an effect on the primary phenotype. The reverse situation will not be treated due to reverse causality challenges (Monsees et al., 2009). In our motivating case study, the Leiden Longevity study (LLS, Houwing-Duistermaat et al. (2009)) families with at least two long-lived siblings are recruited. Obviously, these families do not represent a random sample from the population and inferences cannot be generalized to the whole population, unless the sampling mechanism is properly modelled. Several datasets are measured in the offspring of the long-lived siblings, namely lipidomics, glycomics, metabolomics, and imaging. These offspring share a part of their genetic variation with the long-lived parent and therefore are expected to represent a healthy subpopulation while the partners represent the population. As data example we will model the effect of genetic factors on the secondary traits glucose and triglyceride levels in the offspring (cases) and their partners (controls). To be able to extrapolate results to the general population, we need to account for the over sampling of long-lived subjects in the families of the LLS. There are several multiple-case family studies. For human longevity, GEHA (Genetics of Healthy aging, Skytthe et al. (2011)) used the same study design as the LLS. Other examples are Genetics in Familial Thrombosis (GIFT with at least two cases with thrombosis) (de Visser et al., 2013; Tsonaka et al., 2013) and the ongoing study from Leiden Family Lab (famlab: https://www.leidenfamilylab.nl) which recruits families with at least two cases with social anxiety disorder. The novel methods presented in this paper will also be essential for

modelling secondary phenotypes in these studies.

In the context of case-control studies Monsees et al. (2009) showed that bias can occur when estimating the SNP effect on secondary phenotypes if the primary and secondary phenotypes are associated. This is often the case because both outcomes are measured on the same subjects and secondary phenotypes are typically chosen for their potential associations with the primary phenotype. They also showed that the amount of bias is dependent on the prevalence of the primary phenotype, the strength of the association between the primary and secondary phenotypes, and the association between the tested marker and the primary trait (see Figure 2.1).

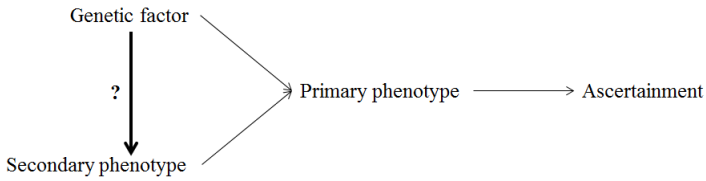

Figure 2.1: Directed acyclic graph representing the case where bias is expected when estimating the association between the genetic marker and the secondary phenotype. Arrows represent existing association between each node of the graph. A secondary phenotype analysis investigates whether there is an association between the genetic factor and the secondary phenotype

To deal with the bias problem, investigators first used ad hoc methods i.e. using controls only, cases only, combined data of cases and controls or joint analysis of cases and controls adjusting for the case-control status. However, several authors showed that these simple approaches can lead to false positive results (Monsees et al., 2009; Lee et al., 1997; Lin and Zeng, 2009). This is due to the sampling design, namely, the secondary phenotype data are not sampled according to the case-control design as the primary phenotype. Several sophisticated methodologies have been developed to correct for the sampling mechanisms and provide unbiased genetic effect estimates: (i) inverse-probability-of-sampling-weighting approaches (Monsees et al., 2009; Richardson et al., 2007; Schifano et al., 2013) which correct for the sampling mechanism by weighting appropriately individuals in case-control studies, (ii) retrospective likelihood-based approaches which indirectly adjust for ascertainment (Lin and Zeng, 2009; He et al., 2011), and (iii) a weighted combination of two estimates obtained with the retrospective likelihood approach in the presence or not of an interaction between SNPs and primary phenotypes (Li and H., 2012).

Even though these approaches can successfully correct for the biased design used to collect the data, they are not directly applicable to more complex designs such as the LLS which motivates this work. In particular, inverse probability weighting approaches require knowledge of the sampling weights for each family. These weights are not available for the LLS because it is unknown what the prevalence of families with at least two nonagenarians is in the population. In addition, the correlations between the family members

cannot be ignored and therefore it is evident that statistical methodology for proper secondary phenotypes analysis in this context is needed. To this end, under the retrospective likelihood framework, we develop a multivariate probit regression model inspired by the work of Najita et al. (2009) to model jointly the distribution of the primary and secondary phenotype. This approach allows us to deal with the ascertainment issue while taking into account the individual relatedness and the genetic and environmental variations.

The paper is organised as follows: in Section 2, we present the retrospective likelihood approach to correct for the over sampling of long-lived subjects and the multivariate probit regression model for the joint modelling of the mixed type primary and secondary phenotypes. In Section 3, we evaluate empirically the performance of the method in terms of bias and efficiency and contrast it with the naive approach which ignores the sampling mechanism. Finally, in Section 4 we illustrate the potential of our proposed method in the analysis of triglyceride levels and glucose in the LLS.

2.2 Methods

2.2.1 Retrospective likelihood approach

Let N be the total number of families in the study. For the family i ($i = 1 \dots N$) of size n_i, let Y_i, X_i and G_i be the $n_i \times 1$ vectors for the case-control status, the secondary phenotype and the genotype, respectively. Motivated by the LLS, we will work under the retrospective likelihood approach to correct for the ascertainment of the families. Such an approach is attractive when modelling the ascertainment mechanism is not straightforward, as in the LLS where sampling depends on the previous generation (an example of a pedigree in LLS is shown in Figure 2.2). In fact the retrospective likelihood approach implicitly corrects for the ascertainment mechanism, under the assumption that the ascertainment depends only on the primary phenotype Y. In particular, for the ith family it holds:

$$P\left(X_i, G_i \mid Y_i, Asc\right) = \frac{P\left(Asc \mid Y_i, G_i, X_i\right) P\left(G_i, X_i \mid Y_i\right)}{P\left(Asc \mid Y_i\right)} = P\left(X_i, G_i \mid Y_i\right), \quad (2.1)$$

with Asc the ascertainment process. By applying Bayes rule we obtain:

$$P\left(X_i, G_i \mid Y_i\right) = \frac{P\left(X_i, Y_i \mid G_i\right) P\left(G_i\right)}{P\left(Y_i\right)} = \frac{P\left(X_i, Y_i \mid G_i\right) P\left(G_i\right)}{\sum_{g \in G} P\left(Y_i \mid g\right) P\left(g\right)}. \quad (2.2)$$

To fully specify (2.2) we need to model properly: the conditional joint distribution of the primary and the secondary phenotypes given the genotype P($X_i, Y_i \mid G_i$), the marginal probability of the primary phenotype P($Y_i \mid G_i$), and the genotype probability of the ith family P(G_i). Each one of these elements are described in Sections 2.2.2 and 2.2.3.

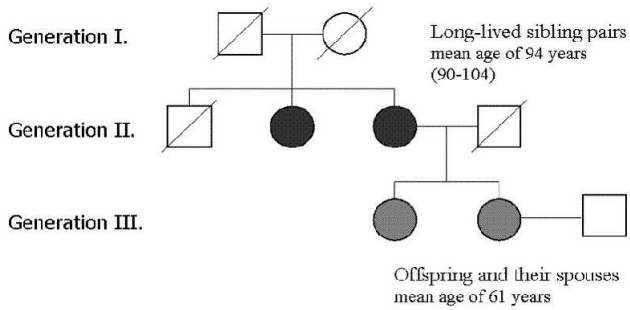

Figure 2.2: Example of a family pedigree from the LLS. Squares and circles represent men and women respectively, crossed symbols represent deceased individuals. In black are the long-lived individuals on whom the ascertainment is based, in grey are the cases of the study (offsprings of long-lived siblings) and in white are the controls.

2.2.2 Mixed-effects models for the analysis of family data

To model the correlation of the phenotypes Y and X within families, a common choice is to use random effects. For the binary primary phenotype we propose to use a multivariate probit model with random effects. The advantage of this model is that it involves only the integrals of the multivariate normal cumulative distribution function for which efficient algorithms have been developed. In contrast, for the more commonly used logistic regression model, the integrals have to be approximated for example by using Gauss-Hermite quadrature which might be computationally intensive for large pedigrees. Let $b_i^Y = \left(b_{i1}^Y, \ldots, b_{in_i}^Y\right)^T$ be a set of family specific random effects designed to handle familial genetic correlation and $G_i = (g_{i1}, \ldots, g_{in_i})^T$ be the vector of genotypes for family i. For the probit model, the observed response Y is viewed as a censored observation from an underlying continuous latent variable Y^* with:

$$Y_{ij} = y_{ij} \Leftrightarrow \gamma_{y_{ij}} < Y_{ij}^* < \gamma_{y_{ij}+1}, Y_{ij} \in \{0,1\}, j = 1,2,...,n_i$$

where $-\infty = \gamma_0 < \gamma_1 < \gamma_2 = +\infty$ are suitable threshold parameters. For the underlying latent variable Y^* we assume the mixed-effects regression model

$$Y_i^* = \alpha_0 + \alpha_1 G_i + \sigma_{G_Y} b_i^Y + \sigma \epsilon_i^Y,$$

where $\epsilon_i^Y \sim N_{n_i}(0, I_{n_i})$ is independent of b_i^Y. Here $\alpha = (\alpha_0, \alpha_1)$ denotes the regression coefficient vector with α_0 the intercept and α_1 the parameter representing the effect of the genotype on Y. At the family level we assume $b_i^Y \sim N_{n_i}(0, \mathbf{R}_i)$, with $\mathbf{R}_i$ the coefficient of relationships matrix with elements $r_{lm} = 2^{-d_{lm}}$ with d_{lm} denoting the genetic distance between subjects l and m in the family. The parameter σ_{G_Y} represents the residual additive genetic variation not explained by g_{ij}. Note that σ_{G_Y} models the

polygenic inheritance in a family.

For identifiability reasons restrictions are required on both the scale and location of Y^*, namely we set $\sigma^2 = 1$ and $\gamma_1 = 0$. Thus, in the mixed-effects probit regression the disease risk $\pi_{ij} = P(Y_{ij} = 1 \mid b_{ij}^Y, g_{ij})$ conditional on the random-effects b_{ij}^Y and genotypic information g_{ij} is modelled as follows

$$P\left(Y_{ij} = 1 \mid g_{ij}, b_{ij}^Y\right) = \Phi\left(\alpha_0 + \alpha_1 g_{ij} + \sigma_{G_Y} b_{ij}^Y\right), \tag{2.3}$$

with $\Phi(z)$ the cumulative distribution function of the standard normal distribution. The marginal density under the probit model takes the form:

$$f(y_{ij} \mid g_{ij}; \alpha, \sigma_b) = \int_{b_i^Y} \int_{\gamma_{y_{ij}}}^{\gamma_{y_{ij}}+1} f(y_{ij}^* \mid g_{ij}, b_i^Y; \alpha, \sigma_b) f(b_i^Y) dy_{ij}^* db_i^Y .$$

To model the secondary phenotype X_i we use a linear mixed model:

$$X_i = \beta_0 + \beta_1 G_i + \sigma_{G_X} b_i^X + \sigma_\epsilon \epsilon_i^X, \tag{2.4}$$

where $\beta = (\beta_0, \beta_1)$ denotes the regression coefficient vector with β_0 the intercept and β_1 the parameter representing the effect of the genotype on X, $b_i^X \sim N_{n_i}(0, \mathbf{R}_i)$ is the random parameter used to model the genetic correlation structure within each family for the secondary trait, and σ_ϵ is the residual standard deviation.

To model jointly X and Y using the model specifications (2.3 and 2.4), we introduce a shared random effect $u_{ij} \sim N(0, 1)$ and propose the following model:

$$\begin{aligned} Y_i^* &= \alpha_0 + \alpha_1 G_i + \sigma_{G_Y} b_i^Y + \sigma_u u_i + \epsilon_i^Y, \\ X_i &= \beta_0 + \beta_1 G_i + \sigma_{G_X} b_i^X + \delta \sigma_u u_i + \sigma_\epsilon \epsilon_i^X, \end{aligned} \tag{2.5}$$

where u_i is assumed to be independent of $b_i^Y, b_i^X, \epsilon_i^Y$, and ϵ_i^X. We introduce a coefficient δ in order to have different phenotypic variances for the random effect u_i. In case of small datasets or small family sizes, it can be better to constrain δ to be equal to 1 for a simpler model. Let Σ_{X_i} and $\Sigma_{Y_i^*}$ denote the corresponding variance-covariance matrices of the marginal distributions of X_i and Y_i^* and let $\Sigma_{XY_i^*}$ be their covariance. The joint distribution of Y^* and X is then $(Y_i^*, X_i) \backsim \mathcal{N}_{2n_i}\left(\begin{bmatrix} \alpha_0 + \alpha_1 G_i \\ \beta_0 + \beta_1 G_i \end{bmatrix}, \begin{bmatrix} \Sigma_{Y_i^*} & \Sigma_{XY_i^*} \\ \Sigma_{XY_i^*} & \Sigma_{X_i} \end{bmatrix}\right)$. In the special case for $n_i = 2$, the variance-covariance matrix becomes:

$$\Sigma_i = \begin{pmatrix} \sigma_{G_Y}^2 + \sigma_u^2 + 1 & \sigma_{G_Y}^2 2^{-d(1,2)} & \sigma_{G_X}\sigma_{G_Y} + \delta\sigma_u^2 & \sigma_{G_X}\sigma_{G_Y} 2^{-d(1,2)} \\ \sigma_{G_Y}^2 2^{-d(1,2)} & \sigma_{G_Y} + \sigma_u^2 + 1 & \sigma_{G_X}\sigma_{G_Y} 2^{-d(1,2)} & \sigma_{G_X}\sigma_{G_Y} + \delta\sigma_u^2 \\ \sigma_{G_X}\sigma_{G_Y} + \delta\sigma_u^2 & \sigma_{G_X}\sigma_{G_Y} 2^{-d(1,2)} & \sigma_{G_X}^2 + \delta^2\sigma_u^2 + \sigma_\epsilon^2 & \sigma_{G_X}^2 2^{-d(1,2)} \\ \sigma_{G_X}\sigma_{G_Y} 2^{-d(1,2)} & \sigma_{G_X}\sigma_{G_Y} + \delta\sigma_u^2 & \sigma_{G_X}^2 2^{-d(1,2)} & \sigma_{G_X}^2 + \delta^2\sigma_u^2 + \sigma_\epsilon^2 \end{pmatrix}. \tag{2.6}$$

Using the properties of the multivariate normal distribution, the joint distribution for the observed primary and secondary phenotypes takes the form:

$$\begin{aligned} P\left(Y_i, X_i \mid G_i\right) &= \int P\left(Y_i^*, X_i \mid G_i\right) dy_i^* \\ &= \int P\left(Y_i^* \mid X_i, G_i\right) P\left(X_i \mid G_i\right) dy_i^* \\ &= P\left(X_i \mid G_i\right) \int P\left(Y_i^* \mid X_i, G_i\right) dy_i^*. \end{aligned}$$

Thus by using the probit regression model for the primary trait we have developed an efficient approach to model the correlation between the primary and secondary trait.

From model (2.5) and the variance-covariance matrix (2.6), several marginal correlations between and within family members can be deduced:

$$\begin{aligned} cor\left(X_{ij}, X_{ij'}\right) &= \frac{\sigma_{G_X}^2 2^{-d\left(j,j'\right)}}{\left(\sigma_{G_X}^2 + \delta^2\sigma_u^2 + \sigma_\epsilon^2\right)} = \rho_X \\ cor\left(Y_{ij}^*, Y_{ij'}^*\right) &= \frac{2^{-d\left(j,j'\right)}\sigma_{G_Y}^2}{\left(\sigma_{G_Y}^2 + \sigma_u^2 + 1\right)} = \rho_Y \\ cor\left(X_{ij}, Y_{ij}^*\right) &= \frac{\sigma_{G_X}\sigma_{G_Y} + \delta\sigma_u^2}{\sqrt{\left(\sigma_{G_X}^2 + \delta^2\sigma_u^2 + \sigma_\epsilon^2\right)\left(\sigma_{G_Y}^2 + \sigma_u^2 + 1\right)}} = \rho_{XY} \\ cor\left(X_{ij}, Y_{ij'}^*\right) &= \frac{2^{-d\left(j,j'\right)}\sigma_{G_X}\sigma_{G_Y}}{\sqrt{\left(\sigma_{G_X}^2 + \delta^2\sigma_u^2 + \sigma_\epsilon^2\right)\left(\sigma_{G_Y}^2 + \sigma_u^2 + 1\right)}} = \rho'_{XY}, \end{aligned}$$

where ρ_{XY} represents the association between the primary and secondary phenotype. We can also derive the closed form for the heritability estimates of the secondary phenotype which quantifies the percentage of genetic variation in the total variance:

$$H^2 = \frac{\sigma_{G_X}^2}{\left(\sigma_{G_X}^2 + \delta\sigma_u^2 + \sigma_\epsilon^2\right)}. \tag{2.7}$$

Note that when genetic factors are included in the model formula (3.2) gives the residual heritability.

2.2.3 Genotype probability

Finally another key component in the formulation of the retrospective likelihood (2.2) is the computation of the genotype probability for each family i. Let G_{mj} and G_{pj} denote the genotypes of the mother and father of an individual j if this individual is a nonfounder member of family i. Under the assumption of random mating and mendelian inheritance, the genotype probabilities can be written as presented by Thomas (2004):

$$P(G_i) = \prod_{j=1}^{J} \begin{cases} P(g_{ij} \mid G_{mj}, G_{pj}) & \text{if } j \text{ is a nonfounder} \\ P(g_{ij}) & \text{if } j \text{ is a founder} \end{cases}.$$

The probabilities $P(g_{ij} \mid G_{pj}, G_{mj})$ are the transmission probabilities which can be modelled using mendelian inheritance. Finally $P(G_{pi})$, $P(G_{mi})$, and $P(g_{ij})$ can be modelled by assuming Hardy-Weinberg proportions $(1-q)^2$, $2q(1-q)$, q^2 which depend on q, the minor allele frequency. Here we propose to use external information for q or to estimate q from the control sample before maximizing the likelihood. Note that when genotypes of the parents are missing the probability can be obtained by summing over the possible parental genotypes. In case of more complex pedigree a recursive algorithm known as peeling (Elston and Stewart, 2013) can be used . For the LLS where families are sibships the probability is as follows:

$$L(\theta; Y, X) = \prod_i \frac{\{P(X_i \mid G_i) \int P(Y_i^* \mid X_i, G_i)\, dy_i^*\} \sum_{G_p} \sum_{G_m} \prod_j P(G_{ij} \mid G_m, G_p) P(G_p) P(G_m)}{\sum_g \sum_{G_p} \sum_{G_m} \int P(Y_i^* \mid g) P(g \mid G_m, G_p) P(G_p) P(G_m)}, \tag{2.8}$$

where $\theta = (\alpha_0, \alpha_1, \sigma_{G_Y}, \beta_0, \beta_1, \sigma_{G_X}, \sigma_\epsilon, \delta, \sigma_u)$ is the model parameters vector.

2.2.4 Estimation and statistical testing

To estimate the parameters of the joint model we maximize the logarithm of the likelihood described in (2.8). This involves a combination of numerical optimization and integration. For the evaluation of the integral in the multivariate normal distribution, we use the deterministic algorithm Miwa described in Miwa et al. (2003). For the optimization, we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm implemented in the function `optim(.)` in R. The BFGS algorithm is a quasi-Newton method, which means that the Hessian matrix does not need to be evaluated directly but is approximated by using specified gradient evaluations. To test for the presence of an effect of the SNPs on the secondary phenotype we use the likelihood ratio test. Note that when the interest of a researcher is solely testing for genetic association a score statistic is an alternative to the likelihood ratio statistic.

2.2.5 Continuous polygenic score

Our approach can also be applied in the case of modelling the association between continuous covariates and secondary phenotypes. For example polygenic scores have been used to summarise genetic effects among an ensemble of SNPs that have been identified in large GWASes (International Schizophrenia Consortium et al., 2009; (IMSGC) et al., 2010; Simonson et al., 2011). Polygenic scores are typically linear combinations of SNPs: $G = \sum_k \delta_k SNP_k$, where $\delta_k = 1$ or δ_k is obtained from previous GWASes. For genetic scores, we need to integrate over the distribution of the polygenic score instead of summing over the genotypes in the denominator of (2.2). For the distribution of the polygenic score we use a multivariate normal distribution $G_i \backsim \mathcal{N}_{n_i}(\mu_g, \sigma_g R_i)$, with μ_g

the mean value of the genetic score, σ_g the standard deviation of the genetic score and R_i the relationship matrix of family i. The likelihood contribution for family i is given by:

$$\frac{P\left(Y_i, X_i \mid G_i\right) P\left(G_i\right)}{P(Y_i)} = \frac{P\left(Y_i, X_i \mid G_i\right) P\left(G_i\right)}{\int_{y_i^*} P(y_i^*) dy_i^*} = \frac{P\left(Y_i, X_i \mid G_i\right) P\left(G_i\right)}{\int_{y^*} \int_{g_i} P(y_i^* \mid g_i) P(g) dy_i^* dg_i}.$$

Computation of the integral $\int_{y^*} \int_g P(y^* \mid g) P(g) dy_* dg$ can be quite intensive and challenging. In order to gain efficiency we write the marginal model of Y^* (2.5) as $Y_i^* = \alpha_0 + b_i^{Y*} + u_i + \epsilon_i^Y$, with $b_i^{Y*} = \sigma_{G_Y} b_i^Y + \alpha_1 G_i$. Now Y_i^* follows the following multivariate normal distribution: $Y_i^* \backsim \mathcal{N}_{n_i}\left(\alpha_0 + \alpha_1 \mu_g, \Sigma_{Y_i^*} + \alpha_1^2 \sigma_g^2 R_i\right)$. Note that when a polygenic risk score is included in the model for the secondary phenotype, the parameter σ_{G_Y} represents the residual polygenic inheritance.

2.2.6 Inclusion of covariates in the model

Often, researchers want to adjust for covariates such as age, sex, treatment etc in the model. Let Z be such a covariate. To estimate the effect Z on the secondary phenotype we propose to maximize the joint likelihood of X and G conditionally on the primary phenotype Y and Z. Thereby we avoid modeling of the distribution of Z within the families. Indeed, under the assumption of independence between genotype and Z we obtain:

$$\begin{aligned} P\left(X_i, G_i \mid Y_i, Z_i\right) = \frac{P\left(X_i, Y_i, Z_i, G_i\right)}{P\left(Y_i, Z_i\right)} &= \frac{P\left(X_i, Y_i \mid G_i, Z_i\right) P\left(G_i\right) P\left(Z_i\right)}{P\left(Y_i | Z_i\right) P\left(Z_i\right)} \\ &= \frac{P\left(X_i, Y_i \mid G_i, Z_i\right) P\left(G_i\right)}{P\left(Y_i | Z_i\right)}. \end{aligned} \tag{2.9}$$

2.3 Simulation Study

A simulation study was set up to evaluate the performance of our proposed method for the estimation of the association between a genetic factor and the secondary phenotype and the estimation of the heritability of the secondary phenotype. We compare the proposed method with the naive approach which is typically followed in practice, namely analysis of the secondary trait without correcting for the sampling mechanism. In particular, in this case, we fit the standard linear mixed-effects model for the secondary phenotype and explicitly model the familial relationships as described in (2.4). The two methods are compared in terms of bias, Root Mean Square Error (RMSE) and 95% coverage probabilities. We consider SNPs (discrete variables) and polygenic scores (continuous variables). Several settings are considered for the disease prevalence, the strength of the association between the genetic factor and the primary phenotype, the strength of the ascertainment mechanism and the number of sibships. We simulated sibships of size 5.

With respect to the familial relationships, we consider only sibships such that our simulation resembles the LLS design. For the prevalence of the primary phenotype we consider two settings namely a disease prevalence of 1% which corresponds to $\alpha_0 \approx -2.32$ and of 5% which corresponds to $\alpha_0 \approx -1.64$. In addition the variance parameters have been chosen such that they correspond to a heritability of 50%. Specifically we use σ_{G_X}=2, $\sigma_{G_Y} = \sqrt{3}$, $\sigma_{u_X} = \sigma_{u_Y} = \sqrt{2}$ and $\sigma_\epsilon = \sqrt{2}$. This corresponds to a correlation of 0.78 between the primary and the secondary phenotypes. To speed up computations, we assume that $\sigma_{u_X} = \sigma_{u_Y}$ when fitting the models to the simulated datasets. For each scenario, 500 datasets are simulated using model (2.5).

2.3.1 Simulation results for a SNP

The genotypes of the SNPs are simulated assuming a minor allele frequency of 0.3 in the population. For the secondary phenotype model the following fixed effects values are used: $\beta_0 = 3.5$ and $\beta_1 = 0.2$, whereas for the primary phenotype model the effect sizes are $\alpha_1 = 0.1$ or 0.5. Finally, for each of the four scenarios (rare or common disease, and weak and strong SNP effect on the primary phenotype) we consider two ascertainment mechanisms, namely the sampled sibships of size five have at least one affected or at least two affected members.

Figure 3.3 presents the estimates and 95% confidence intervals for the scenario of 400 sibships. Figure 3.3 shows that ignoring the sampling mechanism (naive method) leads to biased estimates of the SNP effect and the size of this bias increases with the strength of the ascertainment mechanism and the association between the SNP and the primary phenotype. Overall we observe that the proposed method gives unbiased estimates of the SNP effect on the secondary phenotype. The coverage probabilities reach the nominal level (see section A of supplementary material). Regarding the prevalence of the primary phenotype, we observe that for the naive method bias increases with lower prevalence, while the proposed method remains robust to the lower amount of information due to the rare primary phenotype. In general, the proposed method leads to smaller RMSE than the naive approach and better coverage probabilities.

In Table 2.1 we present the heritability estimates of the secondary phenotype for a common disease, under the various ascertainment mechanisms and the two values of α_1. It is obvious that the heritability estimates are influenced by the ascertainment mechanisms when using the naive approach. Indeed the naive method tends to underestimate the heritability for each mechanism and this underestimation increases as the ascertainment mechanisms become more stringent. The heritability estimates are 25-27% for sibships with at least one affected sibling and drop to 13-14% for sibships with at least 2 affected siblings. On the contrary, the proposed method is robust to the stringency of the ascertainment mechanism.

Next, we study the robustness of our approach to one violation of the model assumptions, namely we simulated under a logit link for the primary phenotype and used the probit link for modelling. Results for the SNP effect and the heritability are presented

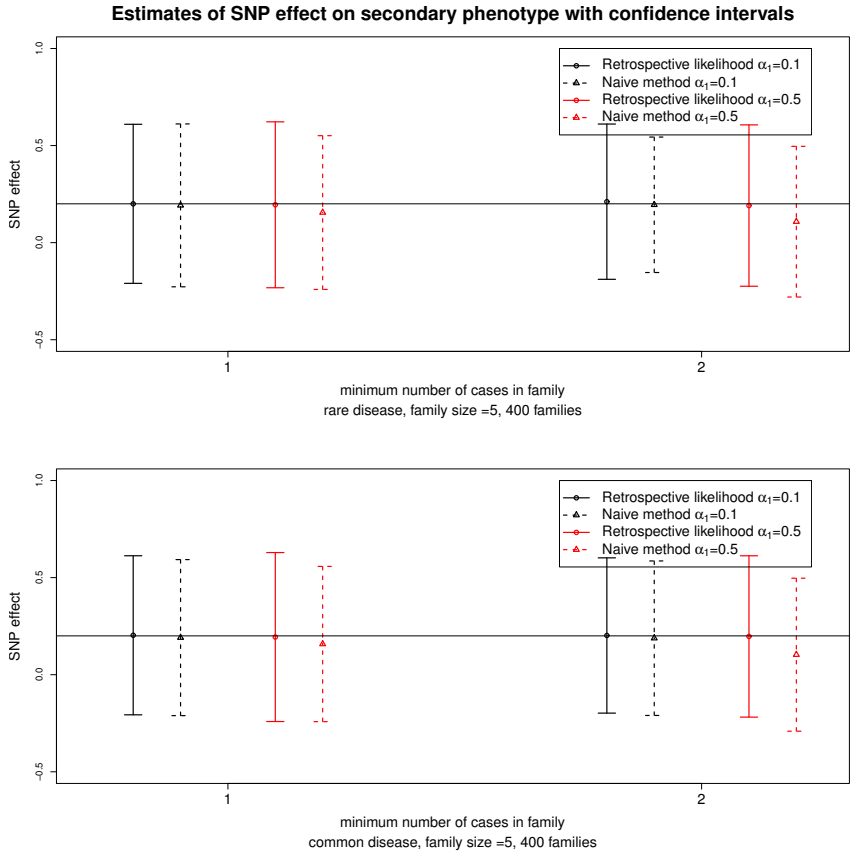

Figure 2.3: Estimates and 95% confidence intervals for the SNP effect on the secondary phenotype for the retrospective likelihood approach and the naive method. Results are obtained from 500 simulated datasets of 400 families for 2 ascertainment schedules. The top and bottom panel correspond to a rare or common primary phenotype with a prevalence around 1% and 5% respectively. In black and red are represented results for small (α_1=0.1) and large (α_1=0.5) effect sizes of the SNP on the primary phenotype, respectively. The horizontal line corresponds to the true SNP effect on the secondary phenotype.

in Table 2.2. These results show that even though our approach gives biased estimates for the primary phenotype model, the parameters estimates for the secondary phenotype model are not affected. All the results are presented in Section A of the Supplementary Material.

Although we focus on parameter estimation, model fitting, and heritability estimation for genetic association with a secondary phenotype, we also investigate the performance of the likelihood ratio test under the null hypothesis of no genetic association with a secondary phenotype at two levels of genetic association with the primary phenotype. In

		SNP model		Polygenic score model	
Ascertainment	α_1	Retrospective	Naive	Retrospective	Naive
1. 2 cases					
	0.10	0.48(0.07)(0.22)	0.13(0.07)(0.37)	0.50(0.03)(0.13)	0.14(0.03)(0.36)
	0.50	0.48(0.07)(0.22)	0.14(0.07)(0.36)	0.52(0.03)0.12)	0.15(0.03)(0.34)
2. 1 case					
	0.10	0.50(0.08)(0.17)	0.25(0.08)(0.25)	0.48(0.04)(0.12)	0.25(0.03)(0.24)
	0.50	0.50(0.08)(0.17)	0.27(0.08)(0.24)	0.50(0.04)(0.10)	0.26(0.04)(0.23)

Table 2.1: Heritability results of the simulation studies for a SNP and a polygenic score: Estimates with standard deviations and RMSE (in brackets) for the heritability of the secondary phenotype for a common disease (prevalence $\approx 5\%$), when families with at least one and at least two cases are sampled and for two values of α_1, i.e. SNP or polygenic score effect on primary phenotype. Datasets consist of 400 families of size 5. Results are based on 500 replicates.

Ascertainment	α_1	β_1	heritability
0.True value		0.200	0.500
1.At least 2 cases			
	0.100	0.199(0.104)(0.104)(0.948)	0.509(0.017)(0.110)
	0.500	0.197(0.106)(0.110)(0.945)	0.516(0.014)(0.108)
2.At least 1 case			
	0.100	0.200(0.104)(0.107)(0.961)	0.510(0.012)(0.096)
	0.500	0.199(0.107)(0.111)(0.960)	0.513(0.010)(0.087)

Table 2.2: Robustness: Estimates of the effect size of the SNP on the secondary phenotype (β_1) and heritability of the secondary phenotype are given for a common disease (prevalence $\approx 5\%$), for the two ascertainment mechanisms and two values of α_1. Into brackets are standard deviations, RMSE and coverage probability (for the effect size only). Datasets consist of 400 families of size 5. Results are based on 500 replicates.

	nominal level (α)	Retrospective likelihood	Naive method
At least 2 cases			
α_1 =0.1			
	0.05	0.0509	0.0580
	0.01	0.0118	0.0152
	0.001	0.0017	0.0025
α_1 =0.5			
	0.05	0.0505	0.0878
	0.01	0.0113	0.0222
	0.001	0.0013	0.0043
At least 1 case			
α_1 =0.1			
	0.05	0.0524	0.0514
	0.01	0.0102	0.0098
	0.001	0.0018	0.0014
α_1 =0.5			
	0.05	0.0522	0.0558
	0.01	0.0098	0.0097
	0.001	0.0009	0.0016

Table 2.3: Type I errors rates for testing for association between a genetic marker and a secondary phenotype for four scenarios. Families with at least one and with at least two cases are considered. Two values for the association between the SNP and the primary phenotype namely α_1 = 0.1 and α_1 = 0.5 are used. Datasets consist of 400 families of size 5. Results are based on 10000 replicates.

each of the four considered scenarios, we simulate 10,000 replicates. In Table 3.2 the emprical type I error rates are given for the rare disease scenario (i.e. prevalence 1%). We observe that while our approach preserves the type I error rate at a nominal level, the naive approach has, systematically, an inflated type I error rate. The type I error rate for the naive method increases with stronger ascertainment and larger SNP effect on the primary phenotype.

2.3.2 Simulation results for a polygenic score

To study the performance of the proposed method for polygenic score, we simulated centered and standardized scores. The parameters of the secondary phenotype model were

chosen as for the SNP simulations: $\beta_0 = 3.5$ and $\beta_1 = 0.2$, whereas for the primary phenotype model effect sizes of $\alpha_1 = 0.1$ or 0.5 were used. Figure 3.4 presents the estimates and confidence intervals for datasets with 400 sibships. Our approach provides unbiased estimates of the effect of the polygenic score on the secondary phenotype. In contrast, the naive approach provides biased estimates and the bias increases when the ascertainment process is more stringent or when α_1 is larger.

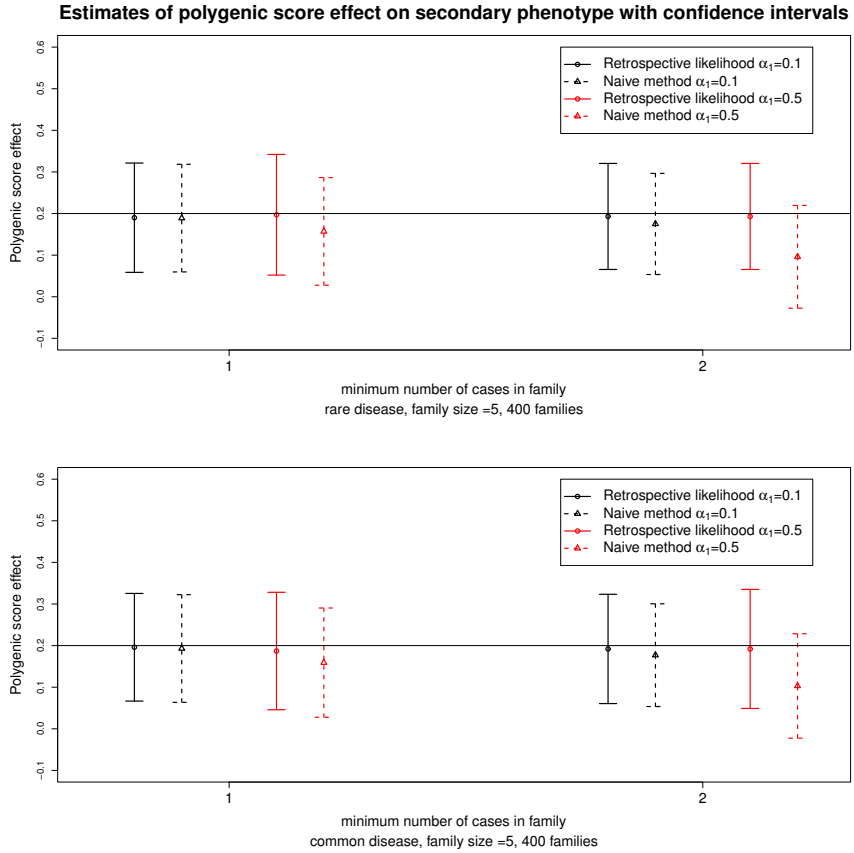

Figure 2.4: Estimates and 95% confidence intervals for the polygenic score effect on the secondary phenotype for the retrospective likelihood approach and the naive method. Results are obtained from 500 simulated datasets of 400 families for 2 ascertainment schedules. The top and bottom panel correspond to a rare or common primary phenotype with a prevalence around 1% and 5% respectively. In black and red are represented results for small (α_1=0.1) and large (α_1=0.5) effect sizes of the polygenic score on the primary phenotype, respectively. The horizontal line corresponds to the true polygenic score effect on the secondary phenotype.

The results of the residual heritability estimates after adjustment for polygenic scores agree with the results obtained when a SNP is included in the model (Table 2.1). The naive

approach did not perform well: estimates between 25-26% and 14-15% for an ascertainment process of at least one affected sibling and at least two affected siblings respectively instead of 50%.

2.4 Application: Analysis of the Leiden Longevity Study

In this Section, we will exemplify our proposed method in the analysis of the LLS briefly introduced in Section 1. The LLS is a family-based study set up to identify mechanisms that contribute to healthy ageing and longevity. The inclusion criteria of the study are sibships with at least two nonagenarian siblings, i.e. the selection takes place at Generation II (Figure 2.2). Several secondary phenotypes and GWAS data have been measured for the offspring of these siblings (Generation III in Figure 2.2) and their partners. Since the offspring have at least one nonagenarian parent, they are also likely to become long-lived. Therefore, the set of offspring and their partners corresponds to a case-control design with related subjects where the offspring in Generation III are considered as cases and their partners as controls. Overall 421 families with 1671 offspring (cases) and 744 partners (controls) have been included in the study. Because the families are relatively small we use the model which assumes an equal variance for the shared effect for the two traits.

Here we model the association between genetic factors and the secondary phenotypes triglyceride and glucose levels. For both traits, there is evidence of an association with human longevity and both traits are normally distributed. For the sake of comparison in addition to our proposed method, we will present results using the naive approach i.e. standard linear mixed model. Analyses using the linear mixed model which conditions also on the case-control status will not be presented because the parameters do not have a comparable interpretation between the two approaches. The p-values presented below are obtained using the likelihood ratio test.

2.4.1 Triglyceride levels analysis

Triglyceride levels have been found to be associated with the primary trait longevity (p-value = 0.0005 for women and p-value = 0.04 for men) and the size of association is sex dependent. Therefore a sex-stratified analysis has been considered further. For the purposes of our illustration, we restricted our analysis to seven genes on chromosome 11 which are known to be associated with Triglyceride levels. These genes are *APOA1, APOA4, APOA5, APOC3, ZNF259, BUD13* and *DSCAML1*. The selection of the genes was performed using the NHGRI-EBI GWAS catalog (Welter et al., 2014). For these genes, we have genotypes of 41 SNPs which have no missing values in our datasets. Triglyceride levels were standardized and we included age as a covariate in the analysis.

We ran the analysis with the constrained approach, i.e. $\delta = 1$. We observe that none of the SNPs analysed is significantly associated with Triglyceride levels either in men or in women, hence for most SNPs the estimates of the effect sizes agree between the

two approaches. The SNPs showing the largest differences are, in men, SNP 22: β_1^{RA} = 0.047 for our Retrospective Approach (RA) and β_1^{NA} = 0.052 for the Naive Approach (NA) and SNP 26: β_1^{RA} = 0.088 and β_1^{NA} = 0.092. For women more SNPs give different estimates between the two approaches, i.e. SNP 1 (β_1^{RA} = 0.024, β_1^{NA} = 0.020), SNP 2 (β_1^{RA} = 7.2e-06 β_1^{NA}=0.006), SNP 13 (β_1^{RA} = -0.013, β_1^{NA} = -0.009) and SNP 19 (β_1^{RA} = 0.011, β_1^{NA} = 0.007) showed the biggest differences. Results for the SNPs are presented in Section B of the Supplementary Material.

We verified whether the assumption of equal variances for the primary and secondary phenotype for the shared effects is justified. We fitted also the model with non constrained δ. We noticed that for some of the SNPs the model parameters are hard to estimate and the estimates of the variances of the shared and residual random effects in the model for the second phenotype are swapped. Overall the estimates of the effect of the SNP on the secondary phenotype are very similar to the model which assumes equal variances. Results of these analyses are presented in Section B of the Supplementary Material.

2.4.2 Glucose levels analysis

In previous analysis of glucose levels in the offspring and partners of the LSS, Mooijaart et al. (2010) studied the association between glucose and a polygenic score. The genetic score was defined as the total number of risk alleles across 15 SNPs which are known to be associated with Type II diabetes. The Generalized Estimating Equation method was applied to take into account the familial relationships. The paper showed that a higher number of Type II diabetes risk alleles is associated with a higher serum concentration of glucose ($p - value$ = 0.016). A statistically significant association was found between glucose level and case-control status (p-value$<$ 0.001). However, the sampling process was not taken into account in the analysis and thus the results might be biased. We applied the proposed method to estimate the heritability of glucose levels and to test for the presence of an association between the glucose levels and the polygenic score. In addition, we applied the naive approach which did not correct for case-control status. We did not stratify according to sex in these analyses.

For this analysis the polygenic score was standardized. Using the Retrospective approach, the association between the genetic score and the glucose level is estimated by $\beta_1^{RA} = 0.630$ with a standard error of $stE = 0.023$ ($p - value = 0.015$). The naive approach also yields a significant association between the genetic score and glucose levels ($\beta_1^{NA} = 0.622$, $stE = 0.026$, $p - value = 0.020$). By using the Naive Approach (NA) we obtained for the glucose levels a genetic variance of $\sigma^2_{G_X} = 0.302$ and a total variance of $\sigma^2_T = 1.322$, which corresponds to a residual heritability of $h^2_{NA} = 0.228$. Our Retrospective approach (RA) yields a genetic variance of $\sigma^2_{G_X} = 0.384$ and a total variance of $\sigma^2_T = 1.457$ which corresponds to a residual heritability of $h^2_{RA} = 0.263$.

2.5 Discussion

In this paper, we developed a new method for the proper analysis of secondary traits for multiple-cases family designs. A key component in our proposed method is the joint modelling of the primary and secondary phenotypes. We developed a multivariate probit model which can also capture the within families dependencies. A retrospective likelihood approach has been followed to correct for the ascertainment process. Thereby unbiased estimates of the association between genetic factors and secondary traits can be obtained. Simulation results showed that our approach preserves the type I error at nominal level and provides accurate estimates irrespective of the disease prevalence, the strength of the association between the genetic variants and the primary phenotype, and the ascertainment mechanism. Another important empirical finding is that the heritability estimates for the secondary traits can be severely underestimated unless the sampling mechanism is taken into account. With respect to the analysis of the motivating case study, for the SNPs the differences between the effect sizes obtained by our proposed method and the naive approach were small. The small differences obtained between the naive and the retrospective approach are mainly due to the small effect sizes of the genetic markers selected on the primary phenotype. Indeed, the three main factors influencing the magnitude of the bias when using the naive approach are the correlation between the secondary phenotype and the primary phenotype, the strength of the ascertainment, and the strength of the association between the genetic marker and the primary phenotype.

Heritability is one of the properties that a trait needs to possess to be declared an endophenotype for a specific disease. The other criteria are: the trait is associated with the disease status in the population, the trait manifests whether illness is active or in remission (state-independent), and the trait and the disease status co-segregate within a family (Gottesman and Gould, 2003). The Leiden Family Lab (https://www.leidenfamilylab.nl) aims to identify endophenotypes for social anxiety disorder. The study comprises families with at least two cases with social anxiety. The methods presented in this paper will be used for the analyses of this study to identify endophenotypes and are relevant for other family studies, as well.

In this paper, we proposed to include additional covariates in the model by using the likelihood conditional on these covariates. Alternatively the joint likelihood of the secondary phenotype, genotype, and covariate conditionally on the primary phenotype can be used. This alternative approach might be more efficient (Balliu et al., 2015). However this likelihood requires distributional assumptions for the covariates within families which can be complex for related individuals. Moreover maximization of the likelihood might become time consuming. Ghosh *et al* (Ghosh et al., 2013) propose a pseudo-likelihood and a profile approach to include covariates in a secondary phenotype analysis for case-control data. This work needs to be extended to family data. A Monte Carlo approach might be considered to compute the integrals (Tsonaka *et al* (Tsonaka et al., 2015)).

Typically there are missing genotypes. In unrelated individuals, genotypes can be imputed based on the haplotype structure obtained from a reference panel. For family data,

the imputation should also take into account the genotypes of other family members. Software exists which can perform such analysis, for example the Genotype Imputation Given Inheritance (GIGI) program (Cheung et al., 2013). However for the computation of the denominator in equation (2.2) these imputed genotype probabilities have to be taken into account.

Due to the computational intensity of the proposed method, it is not yet possible to run full GWAs analyses of secondary phenotypes. However, the proposed method can be used on a set of pre-selected variants e.g. after an initial screening with the naive approach to the primary and secondary phenotypes or when investigating pleiotropic effects. To reduce computation time of the multivariate integrals in the numerator and the denominator, a faster algorithm can be used than the one used in this paper. The randomized Quasi-Monte-Carlo procedure, developed by Genz (1992), is less accurate but faster especially for large pedigrees. Development of less computational intensive methods is one of the topics for future research.

With regard to pleiotropic effects, a criticism of probit random-effects models is that in the presence of high dimensional random effects we cannot move from the subject-specific interpretation for the fixed effects parameters to the population-level interpretation as in the random-intercepts case. When the outcome is binary and families are relatively small, estimation of the intercept and variances terms can be difficult, and consequently coverage probabilities can be poor. Tsonaka et al. (2013) showed efficiency gains by using information on disease prevalence. Their methods need to be adapted to our setting of the analysis of two phenotypes. When the parameters of the primary phenotype model are not of interest and this model is only used to correct for the ascertainment mechanism which is driven by the primary phenotype, we showed that secondary phenotype analyses with the proposed method are robust to using the probit instead of the logit link function.

Future directions in the LSS and Leiden Family Lab Study will address the pending availability of multiple omics and fMRI data, respectively, and joint modelling of several glycans or voxels is of interest. Extending our approach, in this case, is algebraically straightforward, but practical implementation may be challenging due to computational intensity especially with a large number of secondary phenotypes. Use of composite likelihood approaches might be a solution and is our current research topic.

Finally, an attractive alternative approach to properly analyse secondary traits is to apply inverse probability weighting. However, it is crucial to correctly specify the weights. Currently, we do not have sufficient information to be able to estimate these weights for our studies. However with access to electronic records for research, such as information from general practitioners to estimate the weights, it is likely that inverse probability weighting approaches can be developed.