Statistics in Medicine

Heritability Estimation of Dichotomous Phenotypes Using a Liability Threshold Model on Ascertained Family-based Samples

Journal:	*Statistics in Medicine*
Manuscript ID	Draft
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kim, Wonji; Seoul National University, Interdisciplinary Program of Bioinformatics; Brigham and Women's Hospital Department of Medicine, Channing Division of Network Medicine Kwak, Soo Heon; Seoul National University, Department of Internal Medicine Won, Sungho; Seoul National University, Interdisciplinary Program of Bioinformatics; Seoul National University, Department of Public Health Science; Seoul National University, Institute of Health and Environment
Keywords:	Heritability, Liability threshold model, Ascertainment bias

Heritability Estimation of Dichotomous Phenotypes Using a Liability Threshold Model on Ascertained Family-based Samples

Wonji Kim[1,2], Soo Heon Kwak[3] and Sungho Won[1,4,5*]

[1] Interdisciplinary program in Bioinformatics, Seoul National University, Korea

[2] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115, USA

[3] Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Korea

[4] Department of Public Health Sciences, Seoul National University, Seoul, Korea

[5] Institute of Health and Environment, Seoul National University, Seoul, Korea

*To whom correspondence should be addressed.

Running Title

Heritability Estimation of Binary Trait on Ascertained Samples

Abstract

Numerous methods for estimating heritability have been proposed; however, unlike quantitative phenotypes, heritability estimation for dichotomous phenotypes is computationally and statistically complex, and the use of heritability is infrequent. In this study, we developed a statistical method to estimate heritability of dichotomous phenotypes using a Liability Threshold Model in the context of ascertained family-based samples. The Liability Threshold Model assumes dichotomous phenotypes are determined by unobserved latent variables that are normally distributed, and this model can be applied to general pedigree data. The proposed methods were applied to simulated data and Korean type-2 diabetes family-based samples, and the accuracy of estimates provided by the experimental methods was compared with that of established methods.

Keywords

Heritability, Liability threshold model, Ascertainment bias

Summary

We developed a new method for estimating heritability based on the Liability Threshold Model for dichotomous traits which can be applied to the extended pedigree structure. Extensive simulation studies showed that heritability estimates obtained with the proposed methods are generally unbiased even for the ascertained family-based samples. We used the proposed method to estimate the heritability of type-2 diabetes using ascertained family-based samples from Korean families, and those estimates confirmed the practical value of our proposed methods.

Introduction

Phenotypes are affected both by environmental factors and genes, and family members are expected to possess similar phenotypes due to their genetic similarity. Heritability was defined to quantify phenotypic similarity attributable to heritable components, and this concept has been widely used to understand the genetic architecture of phenotypes [1]. For example, heritability can be used to compare the importance of genetic components among different phenotypes. Additionally, if large-scale genetic data are available, genetic correlation matrices can be estimated [2]. These data can then be incorporated into a linear mixed model to provide SNP heritability estimation. SNP heritability provides information regarding the relative proportion of variance attributable to the genotyped SNPs, and this technique can be used to identify the degree of missing heritability.

Estimation of broad-sense heritability requires the study of bilinear relatives such as sibling or monozygotic twins, and in practice, narrow-sense heritability has often been utilized. Narrow-sense heritability is defined as the proportion of the total phenotypic variation explained by additive genetic effects [1]. Various methods have been developed for estimating the heritability of continuous traits. For example, restricted maximum likelihood methods based on the linear mixed model (LMM) [3-5] or polygenic score methods [6] can be used for estimating the heritability of continuous traits. For dichotomous traits, generalized linear mixed models or Liability Threshold Models have been often utilized [7,8]. The Liability Threshold Model assumes there are unobserved continuous liability scores, and subjects are affected if they exceed a certain threshold [9-12].

In this study, we focus on heritability estimation of dichotomous phenotypes. There are multiple factors which can bias variance estimation of dichotomous traits. In particular, family-based samples are typically analyzed using probands. The term proband refers to

instances when family members are brought into a study as a result of other family members already enrolled in the study. Multiple reports indicate that proband analysis can produce substantial bias in variance estimates [4,13,14]. For example, if phenotypes are rare and families are randomly selected, the number of affected individuals is often very small. Therefore families are ascertained through the use of affected probands. In such instances, the majority of the relatives may be unaffected unless the size of the family is very large, and negative correlation can be observed because probands are affected while their relatives are unaffected. Several approaches have been proposed to adjust for such bias. GCTA adjusts estimated heritabilities by assuming that the level of ascertainment bias is same among individuals [4]; however, families are ascertained with probands and the effect of ascertainment bias is heterogeneous according to familial relationship [14]. For example, ascertainment bias for grandparents of the proband is expected to be approximately half that of the parents.

Here, we developed a new method to estimate heritability based on the Liability Threshold Model for binary traits (LTMH) which can be applied to the extended pedigree structure. Using the Expectation-Maximization (EM) algorithm, the proposed method jointly estimates maximum likelihood estimators (MLE) for heritability and coefficients of covariates [15]. Furthermore, the proposed method maximizes the conditional likelihood of disease statuses of probands via a conditional EM (CEM) algorithm [16], and ascertainment bias can be adjusted. We also developed a conditional expected score test (CEST) to determine if heritability is equal to zero. Extensive simulation studies demonstrated that heritability estimates obtained from the proposed methods are generally unbiased even for the ascertained family-based samples. Estimates from GCTA are unbiased for randomly selected families, but the bias turns out to be substantial for ascertained families. Also we found that the CEST for heritability was statistically conservative, but it could achieve reasonable statistical power estimates. Finally, we used the proposed method to estimate the heritability

of type-2 diabetes (T2D) using ascertained family-based samples from Korean families, and those estimates confirmed the practical value of our proposed methods.

Materials and Methods

Notations and Disease Model

We assume that there are n independent families and family i has n_i family members ($i = 1,...,n$). We consider the Liability Threshold Model, and assume dichotomous phenotypes are determined by the unobserved continuous liability score. The liability score of subject j in family i is denoted by L_{ij}, and they are determined by summing the environmental/genetic effects, polygenic effects, and random error. The covariates including environmental/genetic effects for subject j in family i are denoted by $\mathbf{X}_{ij}$, and we assumed that covariates are standardized. In this article, we assumed there are p covariates. The random effects, including polygenic effect and random error for subject j in family i, are denoted by U_{ij}. The vector forms of those components for family i are denoted by:

$$\mathbf{L_i} = \begin{pmatrix} L_{i1} \\ \vdots \\ L_{in_i} \end{pmatrix}, \mathbf{X_i} = \begin{pmatrix} \mathbf{X}_{i1} \\ \vdots \\ \mathbf{X}_{in_i} \end{pmatrix} \text{ and } \mathbf{U_i} = \begin{pmatrix} U_{i1} \\ \vdots \\ U_{in_i} \end{pmatrix}.$$

Liability scores of family members are usually correlated, and we assumed that those are normally distributed as follows:

$$\mathbf{L}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{U}_i, \mathbf{L}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$$

where $\mathbf{U}_i \sim MVN(\mathbf{0}, \boldsymbol{\Sigma}_i)$. We denote $\boldsymbol{\Phi}_i$ to be the kinship coefficient matrix multiplied by two, and $\mathbf{I}_w$ is the $w \times w$ dimensional identity matrix. Under the polygenic model using additivity of genetic effects across loci and linkage equilibrium among loci, we can get:

$$\boldsymbol{\Sigma}_i = \sigma_a^2\boldsymbol{\Phi}_i + \sigma_d^2\mathbf{V}_{di} + \sigma_h^2\mathbf{V}_{hi} + \sigma_{a,d}\mathbf{V}_{adi} + \sigma_e^2\mathbf{I}_{n_i}$$

where σ_a^2, σ_d^2 and σ_e^2 are the variances of additive, dominant, and environmental effects in the population, and σ_h^2 and $\sigma_{a,d}$ are the dominant genetic variance and the covariance of additive and dominant effects in the homozygous population, respectively [17-19]. $\mathbf{V}_{di}$, $\mathbf{V}_{hi}$ and $\mathbf{V}_{adi}$ are the functions of the condensed coefficients of identity [19]. For simplicity, we assume that all variance components other than σ_a^2 and σ_e^2 are zero, and the sum of σ_a^2 and σ_e^2 is equal to one. If we denote heritability as $h^2 = \sigma_a^2/(\sigma_a^2 + \sigma_e^2)$, then the variance-covariance matrix of $\mathbf{\Sigma}_i$ is expressed by

$$\mathbf{\Sigma}_i = h^2\mathbf{\Phi}_i + (1 - h^2)\mathbf{I}_{n_i}.$$

The dichotomous phenotypes for subject *j* in family *i* are denoted by Y_{ij} and these values are coded as 1 for cases and 0 for controls. Phenotype vector for family *i* is denoted by:

$$\mathbf{Y_i} = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{in_i} \end{pmatrix}.$$

In a Liability Threshold Model, Y_{ij} is determined by L_{ij}, and if L_{ij} is larger than a certain threshold value *c*, Y_{ij} becomes 1, and otherwise it becomes 0. *c* can be determined from the prevalence of the diseases as *c* should be the inverse of the cumulative distribution function of the prevalence. For each observed Y_{ij}, we can infer the range of the corresponding L_{ij}, (a_{ij},b_{ij}). For example, if $Y_{ij} = 0$, then L_{ij} is bounded by $(-\infty, c)$, and otherwise, L_{ij} is bounded by (c, ∞). The lower and upper bounds of the liability for the family *i* are denoted by:

$$\mathbf{a_i} = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{in_i} \end{pmatrix} \text{ and } \mathbf{b_i} = \begin{pmatrix} b_{i1} \\ \vdots \\ b_{in_i} \end{pmatrix}.$$

Based on above notations, all subjects can be expressed in the following vector forms:

$$\mathbf{L}=\begin{pmatrix}\mathbf{L}_1\\ \vdots\\ \mathbf{L}_n\end{pmatrix},\mathbf{X}=\begin{pmatrix}\mathbf{X}_1\\ \vdots\\ \mathbf{X}_n\end{pmatrix},\mathbf{U}=\begin{pmatrix}\mathbf{U}_1\\ \vdots\\ \mathbf{U}_n\end{pmatrix},\mathbf{Y}=\begin{pmatrix}\mathbf{Y}_1\\ \vdots\\ \mathbf{Y}_n\end{pmatrix},\mathbf{a}=\begin{pmatrix}\mathbf{a}_1\\ \vdots\\ \mathbf{a}_n\end{pmatrix}\text{ and }\mathbf{b}=\begin{pmatrix}\mathbf{b}_1\\ \vdots\\ \mathbf{b}_n\end{pmatrix}.$$

Under those notations, we assumed that $\mathbf{L}$ follows multivariate normal distribution with mean $\mathbf{X\beta}$ and variance-covariance matrix $\mathbf{\Sigma}$ which exist in a block diagonal matrix consisting of $\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_n$.

Heritability Estimation using the EM Algorithm

The EM (Expectation-Maximization) algorithm [15] was used to estimate h^2 based on the complete data consisting of observed phenotypes, $\mathbf{Y}$, and unobserved liabilities, $\mathbf{L}$. The joint probability density function (pdf) of the complete data can be decomposed into the marginal pdf of $\mathbf{L}$ and the conditional pdf of $\mathbf{Y}$ given that $\mathbf{L}$ has the support of $(\mathbf{a}, \mathbf{b})$. This can be formulated as:

$$f(\mathbf{Y},\mathbf{L})=f(\mathbf{Y}|\mathbf{L})f(\mathbf{L})=f(\mathbf{L})I(\mathbf{a}<\mathbf{L}<\mathbf{b}).$$

If we define the parameters of interest as $\boldsymbol{\theta}=(\boldsymbol{\beta}^t,h^2)^t$, then the log-likelihood of the complete data will be the sum of the log-likelihoods for each family as follows:

$$l(\boldsymbol{\theta};\mathbf{Y},\mathbf{L})=\sum_{i=1}^{n}\left[-\frac{n_i}{2}\log 2\pi-\frac{1}{2}\log|\mathbf{\Sigma}_i|-\frac{1}{2}(\mathbf{L}_i-\mathbf{X}_i\boldsymbol{\beta})^t\mathbf{\Sigma}_i^{-1}(\mathbf{L}_i-\mathbf{X}_i\boldsymbol{\beta})\right].$$

In the E-step of the EM algorithm, the conditional expectation of $\mathbf{L}$ given $\mathbf{Y}$ was taken to the $l(\boldsymbol{\theta};\mathbf{Y},\mathbf{L})$, where the estimates for the parameters of the previous iteration were used. If we assume that the kth iteration has been performed and denote the estimates for the parameters at the kth iteration as $\boldsymbol{\theta}^{(k)}$, then the conditional expectation $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)})$ will be

$$Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)})=E_{\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)}}[l(\boldsymbol{\theta};\mathbf{Y},\mathbf{L})]=\sum_{i=1}^{n}E_{\mathbf{L}_i|\mathbf{Y}_i,\boldsymbol{\theta}^{(k)}}[l_i(\boldsymbol{\theta};\mathbf{Y}_i,\mathbf{L}_i)]=\sum_{i=1}^{n}Q_i(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^{(k)})$$

and

$$Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) = -\frac{n_i}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma}_i| - \frac{1}{2}\left[\mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}\mathbf{A}_i^{(k)}\right) - 2\boldsymbol{\beta}^t\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{B}_i^{(k)} + \boldsymbol{\beta}^t\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\boldsymbol{\beta}\right]$$

where $\mathbf{A}_i^{(k)} = E_{\mathbf{L}_i|\mathbf{Y}_i,\boldsymbol{\theta}^{(k)}}(\mathbf{L}_i\mathbf{L}_i^t)$ and $\mathbf{B}_i^{(k)} = E_{\mathbf{L}_i|\mathbf{Y}_i,\boldsymbol{\theta}^{(k)}}(\mathbf{L}_i)$. $\mathbf{A}_i^{(k)}$ and $\mathbf{B}_i^{(k)}$ are equal to the first moment and the second moment of the multivariate truncated normal, respectively. R package *tmvtnorm* was utilized for calculation [20].

In the M-step of the EM algorithm, we maximize $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$. Since $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ is the concave function, we can find the maximizer by solving for $\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})/\partial \boldsymbol{\theta} = 0$. The partial derivative with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{B}_i^{(k)} - \sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\boldsymbol{\beta} \qquad (1)$$

and, $\boldsymbol{\beta}^{(k)}(h^2)$ which satisfies $\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})/\partial\boldsymbol{\beta} = 0$ becomes

$$\boldsymbol{\beta}^{(k)}(h^2) = \left(\sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{B}_i^{(k)}\right).$$

To emphasize that the root is the function of h^2, it was denoted by $\boldsymbol{\beta}^{(k)}(h^2)$. Unfortunately, there is no closed form of the root in which $\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})/\partial h^2 = 0$, and generalized EM algorithms were applied. $\boldsymbol{\theta}^{(k)}$ was updated using a Newton-Raphson algorithm [21]. After we obtained the maximizer of $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ during the maximization step, we updated $\boldsymbol{\theta}^{(k)}$ to $\boldsymbol{\theta}^{(k+1)}$ and repeated the EM steps until convergence. The detailed algorithm is provided in Appendix (A).

Note that $\hat{\boldsymbol{\beta}}$ is the unbiased estimator of $\boldsymbol{\beta}$ and it can be easily proven by

$$E_{\mathbf{Y}_i}\left(\mathbf{B}_i^{(m)}\right) = E_{\mathbf{Y}_i}\left(E_{\mathbf{L}_i|\mathbf{Y}_i,\boldsymbol{\theta}^{(m)}}(\mathbf{L}_i)\right) = E_{\mathbf{L}_i}(\mathbf{L}_i) = \mathbf{X}_i\boldsymbol{\beta}$$

assuming we obtained $\hat{\boldsymbol{\beta}}$ after m iterations [22].

Lagrangian Multiplier and Karush-Kuhn-Tucker Condition

Unlike $\boldsymbol{\beta}$, the parameter space of h^2 is restricted to $\Theta_{h^2} = \{h^2 : 0 \leq h^2 \leq 1\}$, and the objective function should be maximized under the restriction as follows:

$$\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) \text{ subject to } 0 \leq h^2 \leq 1.$$

This objective function can be maximized using the method of Lagrange multiplier [23] under Karush-Kuhn-Trucker (KKT) conditions [24]. The constraint is equivalent to $-h^2 \leq 0$ and $h^2 - 1 \leq 0$, and by the Lagrangian multiplier, the object function becomes

$$Q^*(\boldsymbol{\theta}, \boldsymbol{\lambda} | \boldsymbol{\theta}^{(k)}) = Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) + \lambda_1 h^2 - \lambda_2 (h^2 - 1)$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^t$. We can find the solution that maximizes $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ subject to $0 \leq h^2 \leq 1$ by finding $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ satisfying the following three conditions known as KKT conditions:

1) *Stationarity* : $\partial Q^*(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\theta}^{(k)}) / \partial \boldsymbol{\theta} = \mathbf{0}$,

2) *Complementary slackness* : $\lambda_1 h^2 = 0$ and $\lambda_2 (1 - h^2) = 0$,

3) *Dual feasibility* : $\lambda_i \geq 0$ for $i = 1, 2$.

More specifically, for the *Stationarity* condition, $\partial Q^*(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\theta}^{(k)}) / \partial \boldsymbol{\beta}$ is identical to $\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)}) / \partial \boldsymbol{\beta}$, providing that $\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(k)}(h^2)$. Replacing $\boldsymbol{\beta}$ with $\boldsymbol{\beta}^{(k)}(h^2)$, we get

$$\left.\frac{\partial Q^*(\boldsymbol{\theta}, \boldsymbol{\lambda} \mid \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}(h^{2*}),\ h^2 = h^{2*}} = \left.\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}(h^{2*}),\ h^2 = h^{2*}} + \lambda_1 - \lambda_2 = 0,$$

and it is equivalent to

$$\left.\frac{\partial Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}(h^{2*}),\ h^2 = h^{2*}} = -\lambda_1 + \lambda_2.$$

Note that to the left of this equation is a function of h^{2*}, denoted by $g^{(k)}(h^{2*})$. Applying *Complementary slackness* conditions to the above equation, $(\lambda_1, \lambda_2, h^2)$ becomes $(0, 0, h^2)$, $(\lambda_1, 0, 0)$, or $(0, \lambda_2, 1)$. If we assume $h^2 = 0$ and $\lambda_2 = 0$, then $g^{(k)}(0) = -\lambda_1$ and it will be non-positive if the assumptions are met by the *Dual feasibility* condition. Similarly, when h^2

$=1$ and $\lambda_1 = 0$ are assumed, $g^{(k)}(1) = \lambda_2$ and it will be non-negative if the assumptions are satisfied. If none of these assumptions are met, λ_1 and λ_2 are automatically zero, and thus optimization can be done without any restrictions on h^2. This concept is illustrated in Figure 1.

Ascertainment Bias-corrected Heritability Estimation

Ascertainment of each family is conducted using probands, and statistical inferences about heritability may be misleading unless ascertainment is correctly adjusted. We assume the first family member in each family is a proband, and the other $n_i - 1$ family members are non-probands. To distinguish probands and non-probands, we added superscripts *P* and *NP*, respectively. Vectors for liabilities, covariates, phenotypes, and bounds of liabilities for non-probands in family *i* are denoted by:

$$\mathbf{L}_i^{NP} = \begin{pmatrix} L_{i2}^{NP} \\ \vdots \\ L_{in_i}^{NP} \end{pmatrix}, \mathbf{X}_i^{NP} = \begin{pmatrix} \mathbf{X}_{i2}^{NP} \\ \vdots \\ \mathbf{X}_{in_i}^{NP} \end{pmatrix}, \mathbf{Y}_i^{NP} = \begin{pmatrix} Y_{i2}^{NP} \\ \vdots \\ Y_{in_i}^{NP} \end{pmatrix}, \mathbf{a}_i^{NP} = \begin{pmatrix} a_{i2}^{NP} \\ \vdots \\ a_{in_i}^{NP} \end{pmatrix} \text{ and } \mathbf{b}_i^{NP} = \begin{pmatrix} b_{i2}^{NP} \\ \vdots \\ b_{in_i}^{NP} \end{pmatrix}.$$

Similarly, those variables pertaining to a proband in family *i* are defined as L_i^P, $\mathbf{X}_i^P$, Y_i^P, a_i^P and b_i^P, respectively. Liability vectors for probands and non-probands across entire families are denoted by:

$$\mathbf{L}^P = \begin{pmatrix} L_1^P \\ \vdots \\ L_n^P \end{pmatrix}, \ \mathbf{L}^{NP} = \begin{pmatrix} \mathbf{L}_1^{NP} \\ \vdots \\ \mathbf{L}_n^{NP} \end{pmatrix} \text{ and } \mathbf{L} = \begin{pmatrix} \mathbf{L}^P \\ \mathbf{L}^{NP} \end{pmatrix},$$

and vectors for other variables are also similarly defined.

To adjust for the effects of ascertainment on heritability estimates, we estimated parameters using the following conditional likelihood:

$$f(\mathbf{Y}^{NP}|\mathbf{Y}^P;\boldsymbol{\theta}) = \frac{f(\mathbf{Y};\boldsymbol{\theta})}{f(\mathbf{Y}^P;\boldsymbol{\theta})}.$$

If we assume $l(\boldsymbol{\theta};\mathbf{Y}) = \log f(\mathbf{Y};\boldsymbol{\theta})$, the log of the conditional likelihood is $l(\boldsymbol{\theta};\mathbf{Y}) - l(\boldsymbol{\theta};\mathbf{Y}^P)$. The objective function of the EM algorithm is a global lower bound for the log-likelihood [25], and if we assume the lower bound $\mathcal{F}(\boldsymbol{\theta})$ for $l(\boldsymbol{\theta};\mathbf{Y})$ and the upper bound $\mathcal{G}(\boldsymbol{\theta})$ for $l(\boldsymbol{\theta};\mathbf{Y}^P)$, then the global lower bound can be obtained by:

$$\log f(\mathbf{Y}^{NP}|\mathbf{Y}^P;\boldsymbol{\theta}) \geq \mathcal{F}(\boldsymbol{\theta}) - \mathcal{G}(\boldsymbol{\theta}).$$

At $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, $\mathcal{F}(\boldsymbol{\theta})$ can be obtained by:

$$\mathcal{F}(\boldsymbol{\theta}) = E_{\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)}}(l(\boldsymbol{\theta};\mathbf{Y},\mathbf{L})) + H\left(f(\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)})\right),$$

where $H(\cdot)$ is the entropy. The upper bound $\mathcal{G}(\boldsymbol{\theta})$ for $l(\boldsymbol{\theta};\mathbf{Y}^P)$ can be defined as $l(\boldsymbol{\theta};\mathbf{Y}^P) + \text{constant}$ [16]. Therefore, the global lower bound of the log-likelihood at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$ becomes:

$$\mathcal{F}(\boldsymbol{\theta}) - \mathcal{G}(\boldsymbol{\theta}) = E_{\mathbf{L}|\mathbf{Y},\boldsymbol{\theta}^{(k)}}(l(\boldsymbol{\theta};\mathbf{Y},\mathbf{L})) - l(\boldsymbol{\theta};\mathbf{Y}^P) + \text{constant}.$$

We assume probands are independent of each other, and proband i was randomly selected from the population with the probability μ_i. Then, $l(\boldsymbol{\theta};\mathbf{Y}^P)$ is simply given by:

$$l(\boldsymbol{\beta};\mathbf{Y}^P) = \sum_{i=1}^{n} l(\boldsymbol{\beta};Y_i^P) = \sum_{i=1}^{n}\left[Y_i^P\alpha_i - \log\left(1+e^{\alpha_i}\right)\right] \text{ where } \alpha_i = \log\frac{\mu_i}{1-\mu_i}.$$

Here μ_i is formulated as a function of the cumulative distribution function of the standard normal, $\Phi(\cdot)$, by:

$$\mu_i = E(Y_i^P) = \Pr\left(Y_i^P = 1\right) = \Pr\left(L_i^P > c\right) = 1 - \Phi(c - \mathbf{X}_i^P\boldsymbol{\beta}).$$

The MLE values for $\boldsymbol{\theta}$ are obtained by iteratively maximizing the objective function until convergence, and the detailed algorithm for maximization is provided in Appendix (B).

Conditional Expected Score Tests

$\boldsymbol{\beta}$ and h^2 are required to parameterize the relationship between covariates and $\mathbf{Y}$ at the unobserved liability scale, and we consider the conditional expected score test (CEST) [15,26,27] because:

$$\frac{\partial l(\boldsymbol{\theta};\mathbf{Y})}{\partial \boldsymbol{\theta}} = E_{\mathbf{L}|\mathbf{Y}}\left[\frac{\partial l(\boldsymbol{\theta};\mathbf{Y},\mathbf{L})}{\partial \boldsymbol{\theta}}\right].$$

For simplicity, we assumed that the prevalence is correctly specified and samples are randomly selected. The conditional expected score based on the complete data for family i is:

$$\mathbf{S}_i = E_{\mathbf{L}|\mathbf{Y}}\begin{bmatrix}\frac{\partial l_i(\boldsymbol{\theta};\mathbf{Y},\mathbf{L})}{\partial \boldsymbol{\beta}} \\ \frac{\partial l_i(\boldsymbol{\theta};\mathbf{Y},\mathbf{L})}{\partial h^2}\end{bmatrix} = \begin{bmatrix}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{B}_i - \mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\boldsymbol{\beta} \\ -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i})\right) - \frac{1}{2}\mathrm{tr}(\mathbf{C}_i\mathbf{A}_i) + \boldsymbol{\beta}^t\mathbf{X}_i^t\mathbf{C}_i\left(\mathbf{B}_i - \frac{1}{2}\mathbf{X}_i\boldsymbol{\beta}\right)\end{bmatrix}$$

where $\mathbf{A}_i = E_{\mathbf{L}|\mathbf{Y}}(\mathbf{L}_i\mathbf{L}_i^t)$, $\mathbf{B}_i = E_{\mathbf{L}|\mathbf{Y}}(\mathbf{L}_i)$ and $\mathbf{C}_i = \partial\boldsymbol{\Sigma}_i^{-1}/\partial h^2$. Note that $\mathbf{A}_i$ and $\mathbf{B}_i$ are also a function of $\boldsymbol{\theta}$. If we assume $\mathbf{S}_{\boldsymbol{\beta}i}$ and S_{h^2i} denote $E_{\mathbf{L}|\mathbf{Y}}[\partial l_i(\boldsymbol{\theta};\mathbf{Y},\mathbf{L})/\partial\boldsymbol{\beta}]$ and $E_{\mathbf{L}|\mathbf{Y}}[\partial l_i(\boldsymbol{\theta};\mathbf{Y},\mathbf{L})/\partial h^2]$, respectively, then the score statistics can be obtained by:

$$\mathbf{S} = \begin{pmatrix}\mathbf{S}_{\boldsymbol{\beta}}^t & S_{h^2}\end{pmatrix}^t \text{ where } \mathbf{S}_{\boldsymbol{\beta}} = \sum_{i=1}^n \mathbf{S}_{\boldsymbol{\beta}i}, \text{ and } S_{h^2} = \sum_{i=1}^n S_{h^2i}.$$

The variance-covariance matrix of $\mathbf{S}$ is calculated using the observed Fisher information matrix [28,29]. The observed Fisher information matrix is given by:

$$\hat{I}(\boldsymbol{\theta}) = \sum_{i=1}^n(\mathbf{S}_i\mathbf{S}_i^t) - \frac{1}{n}\left(\sum_{i=1}^n \mathbf{S}_i\right)\left(\sum_{i=1}^n \mathbf{S}_i^t\right)$$

and it is equivalent to:

$$\hat{I}(\boldsymbol{\theta}) = \begin{pmatrix}\boldsymbol{i}_{\boldsymbol{\beta}} & \boldsymbol{i}_{\boldsymbol{\beta}h^2} \\ \boldsymbol{i}_{h^2\boldsymbol{\beta}} & i_{h^2}\end{pmatrix} = \begin{pmatrix}\sum_{i=1}^n(\mathbf{S}_{\boldsymbol{\beta}i}\mathbf{S}_{\boldsymbol{\beta}i}^t) - \mathbf{S}_{\boldsymbol{\beta}}\mathbf{S}_{\boldsymbol{\beta}}^t/n & \sum_{i=1}^n(\mathbf{S}_{\boldsymbol{\beta}i}S_{h^2i}) - \mathbf{S}_{\boldsymbol{\beta}}S_{h^2}/n \\ \sum_{i=1}^n(S_{h^2i}\mathbf{S}_{\boldsymbol{\beta}i}) - S_{h^2}\mathbf{S}_{\boldsymbol{\beta}}^t/n & \sum_{i=1}^n(S_{h^2i}^2) - S_{h^2}^2/n\end{pmatrix}.$$

Therefore, if we assume p to be the dimension of $\boldsymbol{\beta}$, and $\hat{h^2}$ and $\hat{\boldsymbol{\beta}}$ are MLEs, we can provide the following statistics [28,29]:

$$\mathbf{S}_{\boldsymbol{\beta}}^t\left\{\boldsymbol{i}_{\boldsymbol{\beta}} - \boldsymbol{i}_{\boldsymbol{\beta}\hat{h^2}}\hat{i}_{h^2}^{-1}\boldsymbol{i}_{\hat{h^2}\boldsymbol{\beta}}\right\}^{-1}\mathbf{S}_{\boldsymbol{\beta}} \sim \chi^2(df = p) \text{ under } H_0: \boldsymbol{\beta} = \mathbf{0}.$$

To test if $H_0: h^2 = 0$, the likelihood is maximized at $h^2 = 0$ with 50% probability and at the positive real number at 50% probability under H_0. Thus we consider:

$$\mathbf{S}_{h^2}^{t}\left\{i_{h^2}-\boldsymbol{i}_{h^2\hat{\boldsymbol{\beta}}}\boldsymbol{i}_{\hat{\boldsymbol{\beta}}}^{-1}\boldsymbol{i}_{\hat{\boldsymbol{\beta}}h^2}\right\}^{-1}\mathbf{S}_{h^2}\sim\frac{1}{2}\cdot\mathbf{0}+\frac{1}{2}\cdot\chi^2(df=1)\ \text{ under }\ H_0\text{: }h^2=0.$$

Simulation studies

Simulation studies were conducted under two different scenarios where families were either randomly selected (scenario 1) or ascertained with probands (scenario 2).

For scenario 1, 500 families were randomly generated. For scenario 2, 50,000 families for each replicate were initially generated. Then, 500 probands were selected from affected individuals, and their family members were determined. For both scenarios, we considered nuclear families and the number of siblings at 1, 2, 3 and 4 with proportions of 0.2, 0.3, 0.3 and 0.2, respectively. Liabilities were determined through summation of major genetic effects, polygenic effects, and random errors. Sums of polygenic effects and random errors were generated using multivariate normal distribution with heritability values of 0.05, 0.2 and 0.4. The main genetic effects were obtained using the product of β and the number of disease alleles. Disease allele frequency was assumed to be 0.2, and genotype frequencies were obtained under Hardy Weinberg Equilibrium. Founder genotypes for each family were generated from B (2, 0.2), and non-founder genotypes were obtained by examining Mendelian transmission. β was obtained by h_a^2 and disease allele frequency (p) using the following equation:

$$h_a^2=\frac{2\beta^2p(1-p)}{2\beta^2p(1-p)+1}.$$

h_a^2 was assumed to be 0.005 and β was 0.1253. Once liabilities were generated, they were considered affective if they were larger than the threshold c. Otherwise, they were considered non-affective. c was chosen to maintain the assumed prevalences (q).

The performance of our experimental method was evaluated using 2,000 replicates exhibiting various combinations of heritabilities (h^2) and prevalences (q). For evaluation of

statistical testing of β, the q were set at 0.1 or 0.2, and h^2 was assumed to be 0.2 or 0.4. For evaluation of statistical testing for h^2, we assumed q = 0.05, 0.1 or 0.2 and h^2 = 0, 0.2 and 0.4. All results were compared to GCTA results for each scenario.

Application for Family-based Samples of Type-2 Diabetes

The proposed method was applied to the cross-sectional study of T2D patients conducted by Seoul National University Hospital in Korea. T2D patients were diagnosed according to the World Health Organization criteria for T2D [30]. The study preferentially included T2D patients with a positive family history of T2D in first-degree relatives, and 681 probands were recruited. Family histories of T2D were obtained based on the memory of probands, but the study excluded relatives who were positive for the 75-g oral glucose tolerance test. Subjects of unknown age were also excluded, and 4,149 non-probands, including 1,115 T2D patients and 648 affected probands, remained. For our analyses, the effect of age was adjusted through use as a covariate, and standardized age was incorporated into final analyses. The prevalence of T2D was set at 10.9% [31], and the heritability of T2D was estimated using our experimental method adjusted for ascertainment bias.

Results

Evaluations of simulated samples

We evaluated the accuracy of parameter estimates using simulated data. For scenario 1, we assumed family-based samples were randomly selected, and means and standard deviations (SD) of $\hat{\beta}$ and $\hat{h^2}$ from 2,000 replicates are given in Table 1. The true value of β is assumed to be 0.1253, and estimates for β by LTMH always provide a close approximation of true values. For $\hat{h^2}$, estimates for LTMH and GCTA are similar if the

prevalence is 0.1 or 0.2, although standard errors caused by estimates using LTMH are always smaller than those produced by GCTA. If prevalence is 0.05 and heritability is 0.4, bias of estimates by GCTA becomes much larger. Figure 2 indicates the distribution of $\hat{h^2}$, and both methods accurately estimate high prevalence. Estimates generated by GCTA, however, are more widely distributed than those generated by LTMH, and we can conclude that LTMH provides generally superior performance.

Table 2 provides summaries of parameter estimates for ascertained families. According to the results, the majority of GCTA estimates are 0 and these estimates exhibit ascertainment bias. Estimates of β and h^2 by LTMH, however, are always close to true values and these results show robustness against ascertainment bias (Table 2). Interestingly, standard errors resulting from estimates generated by LTMH analysis of ascertained families are small compared to those observed in the absence of ascertainment. The number of affected individuals is expected to be very small for rare diseases, but ascertainment of affected probands and familial correlations increase the number of affected individuals, which may explain the smaller standard errors observed in heritability estimates of ascertained families. Further investigation, however, is required.

We also evaluated the performance of CEST in the context of hypothesis testing for scenario 1. We assumed $H_0{:}h^2 = 0$, and results detailing empirical sizes are given in Table 3. Our results indicate that LTMH analyses were slightly conservative if q = 0.05 or 0.2, but type-1 error estimates generated by this method are very close to nominal significance levels if q = 0.1. This conservative trend may indicate overestimation of variance. Table 3 also details the statistical power estimates. We assumed that the true h^2 is 0.2 or 0.4, and q is 0.05, 0.1 and 0.2. The statistical power estimates increase as the true heritability, prevalence, or both increase, and large empirical power estimates were obtained in regard to the larger prevalence. We also evaluated the statistical performance of the score tests for β (Tables 4).

Analyses indicate that the score tests for β are not conservative and always preserve the nominal significance level under the null hypothesis, where $H_0{:}\beta = 0$. Empirical power estimates for β were assessed using 2,000 replicates at several significance levels, and these estimates increase as the prevalence, heritability, or both become larger. We also assessed empirical size estimates assuming $H_0{:}h^2 = 0$ for scenario 2 (Table 5). It was more conservative but statistical powers were improved when true h^2 is 0.2 or 0.4 than those for scenario 1.

Applications of LTMH and CEST to Type-2 Diabetes

To evaluate the performance of LTMH using real data, we examined the family-based samples from the T2D dataset. Table 6 shows the descriptive statistics [32]. There were 1,736 T2D patients (36.75%), and average age for entire samples was 48.63 years old with SD of 15.7 . The proportions of males and females are similar. All non-probands are the first-degree relatives of probands, and the familial relationships observed most often are siblings (59.22%) and offspring (32.85%).

LTMH was used to examine the family-based samples derived from the T2D dataset, and heritability of T2D was estimated. Estimated heritability of T2D was 29.44%, and it was statistically significant under the significance level of 0.05 (P-value = 1.20×10^{-5}). This finding is slightly overestimated in comparison to other determinations of heritability estimates for T2D (26%) using the ACE model based on twin data [33]. This difference may be attributable to racial differences. The coefficient estimate for non-standardized age was 0.051 (0.8 for standardized age), which means that the threshold for disease is reduced by 0.051 at the liability scale if age increases by 1. The function of age is well described in Figure 3A, which illustrates the probability of being affected by T2D as a function of age. Results demonstrate that the risk increases monotonically by age, reflecting the reduction effect on

disease threshold. Individuals with a higher number of T2D affected relatives exhibit greater risk. In comparison to random samples, the influence of family history is greater at a young age, and determining familial risk for early-onset T2D is highly important (Figure 3B).

Discussion

In this article, we proposed a new method to estimate the heritability of a dichotomous trait based on the Liability Threshold Model for ascertained family-based samples. A simulation study demonstrated that LTMH generally provides more accurate estimates of heritabilities than does GCTA, and the differences between these methods are substantial in the context of ascertained families. To our knowledge, there is no method to effectively approach ascertained samples to estimate heritability of dichotomous traits. Additionally, we assessed the statistical performance of CEST analysis. Statistical power estimates were evaluated under various experimental conditions, and substantial power improvement was observed in the context of common diseases as opposed to that seen for rare diseases.

Despite the power improvement provided by the proposed methods, there are limitations. First, the CEST for h^2 was conservative. We found that the likelihood for h^2 is not symmetric under the null hypothesis, and this may be attributable to the misspecified weights for 0 and $\chi^2(df = 1)$ for the distribution of the CEST under H_0. Fortunately, we found that such inflation does not affect the statistical power of our analysis, but certain modifications such as bootstrapping are necessary. Second, the proposed method is the computationally intensive when the family size, n_i, is large, and the expected computational time is proportional to $O(\max_i n_i^3)$. The most significant computational burden arises from the calculation of conditional expectation in the E-step of the EM algorithm. The

computational burden can be reduced by reducing the number of iterations for the EM algorithm or by approximating the moment of the multivariate truncated normal. The former can be achieved by using EM acceleration methods which can make EM dramatically faster. These include Aitken acceleration, conjugate gradient acceleration, quasi-Newtonian acceleration, and parameter expansion acceleration [34-38]. For the latter, conditional expectation may be approximated using certain numerical algorithms such as Laplace approximation. Investigation of these techniques will be the focus of future research.

Heritability shows important utility for genetic epidemiology; however, heritability estimation of dichotomous phenotypes can be extremely complicated due to ascertainment bias. Despite several limitations, our proposed method successfully enabled heritability estimation of dichotomous traits in ascertained families, and this method may provide a promising strategy to estimate the narrow-sense heritability of various diseases. LTMH is implemented in R language, and source codes are freely available at http://healthstat.snu.ac.kr/software/LTMH.

Acknowledgement

This work was supported by the Industrial Core Technology Development Program (20000134) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) and by the National research Foundation of Korea (2017M3A9F3046543).

References

1. Visscher PM, Hill WG, Wray NR. Heritability in the genomics era—concepts and misconceptions. *Nature reviews genetics.* 2008;9(4):255.
2. Fedko IO, Hottenga J-J, Medina-Gomez C, et al. Estimation of genetic relationships between individuals across cohorts and platforms: application to childhood height. *Behavior genetics.* 2015;45(5):514-528.
3. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature genetics.* 2010;42(7):565.
4. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics.* 2011;88(1):76-82.
5. Vattikuti S, Guo J, Chow CC. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS genetics.* 2012;8(3):e1002637.
6. Dudbridge F. Power and predictive accuracy of polygenic risk scores. *PLoS genetics.* 2013;9(3):e1003348.
7. Papachristou C, Ober C, Abney M. Genetic variance components estimation for binary traits using multiple related individuals. *Genetic epidemiology.* 2011;35(5):291-302.
8. Burton PR, Tiller KJ, Gurrin LC, Cookson WO, Musk AW, Palmer LJ. Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society.* 1999;17(2):118-140.
9. Dempster ER, Lerner IM. Heritability of threshold characters. *Genetics.* 1950;35(2):212.
10. Van Vleck L. Estimation of heritability of threshold characters. *Journal of Dairy Science.* 1972;55(2):218-225.

11. Hoeschele I, Tier B. Estimation of variance components of threshold characters by marginal posterior modes and means via Gibbs sampling. *Genetics Selection Evolution.* 1995;27(6):519.

12. Lee SH, Wray NR, Goddard ME, Visscher PM. Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics.* 2011;88(3):294-305.

13. Sawyer S. Maximum likelihood estimators for incorrect models, with an application to ascertainment bias for continuous characters. *Theoretical Population Biology.* 1990;38(3):351-366.

14. Park S, Lee S, Lee Y, et al. Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families. *BMC medical genetics.* 2015;16(1):62.

15. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society Series B (methodological).* 1977:1-38.

16. Jebara T, Pentland A. Maximum conditional likelihood via bound maximization and the CEM algorithm. Paper presented at: Advances in neural information processing systems1999.

17. Fisher RA. XV.—The correlation between relatives on the supposition of Mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh.* 1919;52(2):399-433.

18. Abney M, McPeek MS, Ober C. Estimation of variance components of quantitative traits in inbred populations. *The American Journal of Human Genetics.* 2000;66(2):629-650.

19. Jacquard A. *The genetic structure of populations.* Vol 5: Springer Science & Business Media; 2012.

20. Wilhelm S, Manjunath B. tmvtnorm: A package for the truncated multivariate normal distribution. *sigma.* 2010;2(2).

21. Atkinson KE. *An introduction to numerical analysis.* John Wiley & Sons; 2008.

22. Weiss NA. *A course in probability.* Addison-Wesley; 2006.

23. Bertsekas DP. *Constrained optimization and Lagrange multiplier methods.* Academic press; 2014.

24. Kuhn HW, Tucker AW. Nonlinear programming. In: *Traces and emergence of nonlinear programming.* Springer; 2014:247-258.

25. Neal RM, Hinton GE. A view of the EM algorithm that justifies incremental, sparse, and other variants. In: *Learning in graphical models.* Springer; 1998:355-368.

26. Fisher RA. Theory of statistical estimation. Paper presented at: Mathematical Proceedings of the Cambridge Philosophical Society1925.

27. Finkelstein DM, Wang R, Ficociello LH, Schoenfeld DA. A score test for association of a longitudinal marker and an event with missing data. *Biometrics.* 2010;66(3):726-732.

28. Rao CR. Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. Paper presented at: Mathematical Proceedings of the Cambridge Philosophical Society1948.

29. Scott WA. Maximum likelihood estimation using the empirical fisher information matrix. *Journal of Statistical Computation and Simulation.* 2002;72(8):599-611.

30. Organization WH. Definition and diagnosis of diabetes mellitus and intermediate hyperglycaemia: report of a WHO/IDF consultation. 2006.

31. Ko S-H, Han K, Lee Y-h, et al. Past and Current Status of Adult Type 2 Diabetes Mellitus Management in Korea: A National Health Insurance Service Database Analysis. *Diabetes & metabolism journal.* 2018;42(2):93-100.

32. Song YE, Lee S, Park K, Elston RC, Yang H-J, Won S. ONETOOL for the analysis of family-based big data. *Bioinformatics*. 2018;1:3.

33. Poulsen P, Kyvik KO, Vaag A, Beck-Nielsen H. Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance–a population-based twin study. *Diabetologia*. 1999;42(2):139-145.

34. Laird N, Lange N, Stram D. Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*. 1987;82(397):97-105.

35. Jamshidian M, Jennrich RI. Conjugate gradient acceleration of the EM algorithm. *Journal of the American Statistical Association*. 1993;88(421):221-228.

36. Lange K. A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*. 1995:425-437.

37. Lange K. A quasi-Newton acceleration of the EM algorithm. *Statistica sinica*. 1995:1-18.

38. Liu C, Rubin DB, Wu YN. Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika*. 1998;85(4):755-770.

Table 1. Accuracy of $\hat{\beta}$ and $\widehat{h^2}$ from randomly selected families (scenario 1). Parameter estimates from 2,000 replicates were summarized using mean (top) and standard error (bottom). The true value of β is 0.1253. SD is standard deviation.

Heritability	Prevalence	LTMH		GCTA
		β	h^2	h^2
0.05	0.05	0.1226 (0.0223)	0.0933 (0.0971)	0.1105 (0.1303)
	0.1	0.1281 (0.0181)	0.0660 (0.0716)	0.0734 (0.0828)
	0.2	0.1277 (0.016)	0.0584 (0.0538)	0.0563 (0.0539)
0.2	0.05	0.1267 (0.0223)	0.2184 (0.1282)	0.2511 (0.1852)
	0.1	0.1239 (0.0190)	0.1950 (0.0993)	0.2111 (0.1219)
	0.2	0.1285 (0.0164)	0.2106 (0.0725)	0.2115 (0.0775)
0.4	0.05	0.1309 (0.0229)	0.4324 (0.1313)	0.5546 (0.2437)
	0.1	0.1276 (0.0225)	0.4230 (0.1315)	0.4825 (0.1377)
	0.2	0.1286 (0.0189)	0.4181 (0.0950)	0.4486 (0.085)

Table 2. Accuracy of $\hat{\beta}$ and $\hat{h^2}$ from ascertained families (scenario 2) Parameter estimates from 2,000 replicates were summarized using mean (top) and standard error (bottom). The true value of β is 0.1253.

Heritability	Prevalence	LTMH		GCTA
		β	h^2	h^2
0.05	0.05	0.1335 (0.0193)	0.0474 (0.0376)	1.72×10^{-6} (4.47×10^{-7})
	0.1	0.1233 (0.0181)	0.0336 (0.0339)	1.96×10^{-6} (2.01×10^{-7})
	0.2	0.1194 (0.0144)	0.0304 (0.0287)	1.83×10^{-6} (3.78×10^{-7})
0.2	0.05	0.1234 (0.0199)	0.2018 (0.0437)	1.01×10^{-6} (9.18×10^{-8})
	0.1	0.1257 (0.0135)	0.2086 (0.0342)	0 (0)
	0.2	0.1239 (0.0153)	0.1692 (0.0407)	1.01×10^{-6} (7.40×10^{-8})
0.4	0.05	0.1358 (0.0189)	0.4004 (0.0449)	0 (0)
	0.1	0.1167 (0.0144)	0.3868 (0.0339)	0 (0)
	0.2	0.1186 (0.0150)	0.4090 (0.0444)	0 (0)

Table 3. Type-1 error and power estimates of the proposed test for H_0:$h^2 = 0$ under scenario 1. The empirical sizes ($h^2 = 0$) and powers ($h^2 = 0.2$ and 0.4) were estimated using 2,000 replicates at three significance levels. We considered prevalence of 0.05, 0.1, and 0.2.

Heritability	Prevalence	Significance level		
		0.01	0.05	0.1
0	0.05	0.0015	0.0115	0.0285
	0.1	0.0050	0.0480	0.1020
	0.2	0.0015	0.0200	0.0505
0.2	0.05	0.0485	0.2260	0.3990
	0.1	0.3420	0.6730	0.8055
	0.2	0.6210	0.8675	0.9405
0.4	0.05	0.4575	0.8190	0.9050
	0.1	0.9395	0.9930	0.9960
	0.2	1.0000	1.0000	1.0000

Table 4. Type-1 error and power estimates of the proposed test for H_0:$\boldsymbol{\beta} = \mathbf{0}$ under scenario 1. The empirical sizes ($h_a^2 = 0$) and powers ($h_a^2 = 0.005$) were estimated using 2,000 replicates at three significance levels. We considered heritability of 0.2 and 0.4, and prevalence of 0.1 and 0.2.

h_a^2	Heritability	Prevalence	Significance level		
			0.01	0.05	0.1
0	0.2	0.1	0.0155	0.0661	0.1023
		0.2	0.0120	0.0560	0.0900
	0.4	0.1	0.0060	0.0480	0.0940
		0.2	0.0130	0.0580	0.1020
0.005	0.2	0.1	0.1303	0.3372	0.4713
		0.2	0.4460	0.6800	0.7980
	0.4	0.1	0.2740	0.5340	0.6640
		0.2	0.3540	0.6000	0.7180

Table 5. Type-1 error and power estimates of the proposed test for H_0:$h^2 = 0$ under scenario 2. The empirical sizes ($h^2 = 0$) and powers ($h^2 = 0.2$ and 0.4) were estimated using 2,000 replicates at three significance levels. We considered prevalence of 0.05, 0.1, and 0.2.

Heritability	Prevalence	Significance level		
		0.01	0.05	0.1
0	0.05	0.0000	0.0025	0.0100
	0.1	0.0005	0.0045	0.0095
	0.2	0.0000	0.0075	0.0215
0.2	0.05	0.4735	0.8110	0.9185
	0.1	0.8520	0.9660	0.9850
	0.2	0.8155	0.9540	0.9855
0.4	0.05	1.0000	1.0000	1.0000
	0.1	1.0000	1.0000	1.0000
	0.2	1.0000	1.0000	1.0000

Table 6. Demographic characteristics of study participants. For categorical variables, the number of subjects and their proportions are provided. For continuous variables, means and standard deviations are provided.

	Proband	Non-proband
Disease status		
T2D	648 (100%)	1,115 (26.87%)
Normal	0 (0%)	3,034 (73.13%)
Sex		
Male	308 (47.53%)	2,058 (49.6%)
Female	340 (52.47%)	2,091 (50.4%)
Age	55.44 (10.7)	47.56 (16.09)
Relationship of relatives with proband		
Parents		329 (7.93%)
Sibling		2,457 (59.22%)
Offspring		1,363 (32.85%)

Legend of Figures

Figure 1. Illustration of KKT condition using a toy example. The exemplary concave function $Q(h^2)$ was created to enable determination of the optimal value that maximizes $Q(h^2)$ within the parameter space. The parameter h^2 can be between zero and one, and the parameter space for this value is grayed out. (A) If the value that maximizes $Q(h^2)$ is negative, the tangent slopes at both zero and one will be negative. A tangent slope that is negative at one violates the KKT conditions, however, a negative tangent slop at zero satisfies the KKT conditions, so the maximizer within the parameter space is zero. (B) When the value which maximizes $Q(h^2)$ is greater than 1, the optimal value is one since positive tangent slope at one meets the KKT conditions. (C) When the maximizer is located in the parameter space, tangent slopes at both boundaries of the parameter space do not satisfy the KKT conditions. Therefore, restrictions do not affect the result of optimization.

Figure 2. Boxplots for $\hat{h^2}$ for randomly selected families (scenario 1). True heritability was 0.05 (top), 0.2 (middle), and 0.4 (bottom) and was indicated as a gray dashed line.

Figure 3. Estimation of risks for T2D according to age. For a certain individual, we assume that he/she has two parents and one younger sibling, and the risk of T2D development was calculated as a function of his/her age and the number of affected family members. The X-axis indicates age of individual, and the age of his/her father and mother were assumed to be 29 years old. The younger sibling was assumed to be 3 years younger than the participant. h^2 and the coefficient of unstandardized age were set to be 0.2944 and 0.051, respectively. (A) Probability of the participant being affected according to the number of affected family members, and (B) relative risks of being affected according to the number of affected family members.

Appendix (A)

Numerical analysis for optimization of the heritability in M-step of EM algorithm

The first derivative of $Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})$ with respect to h^2 is given by

$$\frac{\partial Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{\partial h^2} = -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i})\right) - \frac{1}{2}\mathrm{tr}\left(\mathbf{C}_i\mathbf{A}_i^{(k)}\right) + \boldsymbol{\beta}^t\mathbf{X}_i^t\mathbf{C}_i\left(\mathbf{B}_i^{(k)} - \frac{1}{2}\mathbf{X}_i\boldsymbol{\beta}\right) \quad (2)$$

where $\mathbf{C}_i = \partial\boldsymbol{\Sigma}_i^{-1}/\partial h^2 = -\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i})\boldsymbol{\Sigma}_i^{-1}$. Then, the objective function becomes

$$\mathcal{M}(h^2) = \sum_{i=1}^{n} \frac{\partial Q_i(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(k)})}{\partial h^2}\bigg|_{\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}(h^2)} = 0.$$

Similarly, we can get the first derivative of $\mathcal{M}(h^2)$ with respect to h^2 as follows,

$\mathcal{M}'(h^2)$

$$= \sum_{i=1}^{n}\Bigg[-\frac{1}{2}\mathrm{tr}\left(\mathbf{C}_i(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i})\right) - \frac{1}{2}\mathrm{tr}\left(\mathbf{H}_i\mathbf{A}_i^{(k)}\right) + \left(\mathbf{X}_i\mathbf{F}^{(k)}\right)^t\mathbf{C}_i\left(\mathbf{B}_i^{(k)} - \frac{1}{2}\mathbf{X}_i\boldsymbol{\beta}^{(k)}(h^2)\right)$$
$$+ \left(\mathbf{X}_i\boldsymbol{\beta}^{(k)}(h^2)\right)^t\mathbf{H}_i\left(\mathbf{B}_i^{(k)} - \frac{1}{2}\mathbf{X}_i\boldsymbol{\beta}^{(k)}(h^2)\right) - \frac{1}{2}\left(\mathbf{X}_i\boldsymbol{\beta}^{(k)}(h^2)\right)^t\mathbf{C}_i\mathbf{X}_i\mathbf{F}^{(k)}\Bigg]$$

where $\mathbf{H}_i = \partial\mathbf{C}_i/\partial h^2 = -2\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i})\mathbf{C}_i$ and

$\mathbf{F}^{(k)}$

$= \partial\boldsymbol{\beta}^{(k)}(h^2)/\partial h^2$

$$= -\left(\sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i^t\mathbf{C}_i\mathbf{X}_i\right)\left(\sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{B}_i^{(k)}\right) + \left(\sum_{i=1}^{n}\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\right)^{-1}$$
$$\left(\sum_{i=1}^{n}\mathbf{X}_i^t\mathbf{C}_i\mathbf{B}_i^{(k)}\right).$$

Finally, h^2 is updated according to the following iterative steps using $\mathbf{A}_i^{(k)}$ and $\mathbf{B}_i^{(k)}$ which were calculated at the previous E-step,

$$h^2_{\text{new}} = h^2_{\text{old}} - \frac{\mathcal{M}(h^2_{\text{old}})}{\mathcal{M}'(h^2_{\text{old}})}.$$

Appendix (B)

Numerical analysis for maximizing the global lower bound

If we denote the global lower bound for the conditional log-likelihood as $\mathbb{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, then the first derivative of $\mathbb{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}$, $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, is given by

$$\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \begin{pmatrix} \dfrac{\partial \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\beta}} + \dfrac{\partial l(\boldsymbol{\beta}; \mathbf{Y}^P)}{\partial \boldsymbol{\beta}} \\ \dfrac{\partial \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial h^2} \end{pmatrix}.$$

Here, it should be noted that $\partial \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})/\partial \boldsymbol{\theta}$ is equivalent to the equations (1) and (2) in the Method and Appendix (A). Using the chain rule, we can easily obtain the first derivative of $l(\boldsymbol{\beta}; \mathbf{Y}^P)$ with respect to $\boldsymbol{\beta}$ as follows,

$$\frac{\partial l(\boldsymbol{\beta}; \mathbf{Y}^P)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left[\frac{\partial l(\boldsymbol{\beta}; Y_i^P) \partial \alpha_i \partial \mu_i}{\partial \alpha_i \quad \partial \mu_i \partial \boldsymbol{\beta}} \right] = \sum_{i=1}^{n} \left[\frac{(Y_i^P - \mu_i)}{\mu_i(1-\mu_i)} \phi(c - \mathbf{X}_i^P \boldsymbol{\beta}) (\mathbf{X}_i^P)^t \right]$$

where $\phi(\cdot)$ is the probability density function for the standard normal. To apply Newton-Raphson algorithm for the objective function $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, we derive the first derivative of $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ with respect to $\boldsymbol{\theta}^t$, $\mathbf{J}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, as follows,

$$\mathbf{J}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) = \begin{pmatrix} \dfrac{\partial^2 \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}} + \dfrac{\partial^2 l(\boldsymbol{\beta}; \mathbf{Y}^P)}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}} & \dfrac{\partial^2 \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial h^2 \partial \boldsymbol{\beta}} \\ \dfrac{\partial^2 \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\beta}^t \partial h^2} & \dfrac{\partial^2 \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial (h^2)^2} \end{pmatrix}$$

and each term is given by

$$\frac{\partial^2 \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}}$$

$$= \sum_{i=1}^{n} \left(\mathbf{X}_i^t \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i\right), \frac{\partial^2 l(\boldsymbol{\beta}; \mathbf{Y}^P)}{\partial \boldsymbol{\beta}^t \partial \boldsymbol{\beta}}$$

1
$$= \sum_{i=1}^{n} \left[\frac{\phi(c - \mathbf{X}_i^P \boldsymbol{\beta})}{\mu_i(1-\mu_i)} \mathbf{X}_i^P \left\{ \frac{(Y_i^P - \mu_i)(2\mu_i - 1)}{\mu_i(1-\mu_i)} \phi(c - \mathbf{X}_i^P \boldsymbol{\beta}) + (Y_i^P - \mu_i)(c - \mathbf{X}_i^P \boldsymbol{\beta}) \right.\right.$$

$$\left.\left. - \phi(c - \mathbf{X}_i^P \boldsymbol{\beta}) \right\} (\mathbf{X}_i^P)^t \right],$$

2 $$\frac{\partial^2 \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial h^2 \partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \left(\mathbf{X}_i^t \mathbf{C}_i \mathbf{B}_i^{(k)}\right) - \sum_{i=1}^{n} \left(\mathbf{X}_i^t \mathbf{C}_i \mathbf{X}_i \boldsymbol{\beta}\right),$$

3 and

4 $$\frac{\partial^2 \mathrm{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})}{\partial (h^2)^2} = -\frac{1}{2}\mathrm{tr}\left(\mathbf{C}_i(\boldsymbol{\Phi}_i - \mathbf{I}_{n_i})\right) - \frac{1}{2}\mathrm{tr}\left(\mathbf{H}_i \mathbf{A}_i^{(k)}\right) + \boldsymbol{\beta}^t \mathbf{X}_i^t \mathbf{H}_i \left(\mathbf{B}_i^{(k)} - \frac{1}{2}\mathbf{X}_i \boldsymbol{\beta}\right).$$

5 With these terms, we iteratively update $\boldsymbol{\theta}$ using the following equation until convergence,

6 $$\boldsymbol{\theta}^{\text{new}} = \boldsymbol{\theta}^{\text{old}} - \mathbf{J}^{-1}(\boldsymbol{\theta}^{\text{old}}|\boldsymbol{\theta}^{(k)})\, \mathcal{H}(\boldsymbol{\theta}^{\text{old}}|\boldsymbol{\theta}^{(k)}).$$

7

8

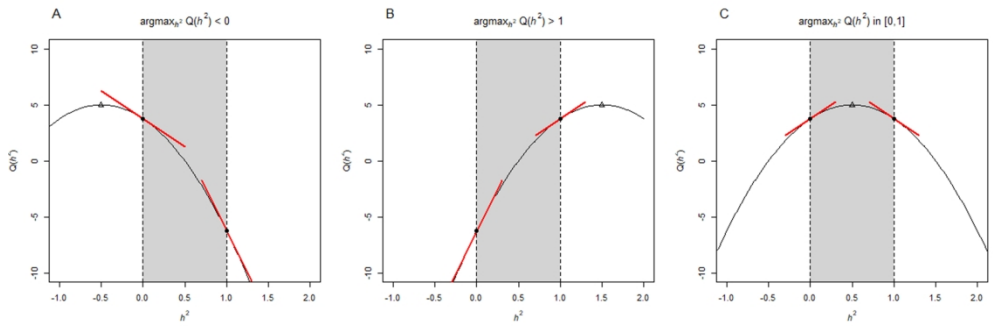

Figure 1. Illustration of KKT condition using a toy example. The exemplary concave function $Q(h^2)$ was created to enable determination of the optimal value that maximizes $Q(h^2)$ within the parameter space. The parameter h^2 can be between zero and one, and the parameter space for this value is grayed out. (A) If the value that maximizes $Q(h^2)$ is negative, the tangent slopes at both zero and one will be negative. A tangent slope that is negative at one violates the KKT conditions, however, a negative tangent slop at zero satisfies the KKT conditions, so the maximizer within the parameter space is zero. (B) When the value which maximizes $Q(h^2)$ is greater than 1, the optimal value is one since positive tangent slope at one meets the KKT conditions. (C) When the maximizer is located in the parameter space, tangent slopes at both boundaries of the parameter space do not satisfy the KKT conditions. Therefore, restrictions do not affect the result of optimization.

330x110mm (236 x 236 DPI)

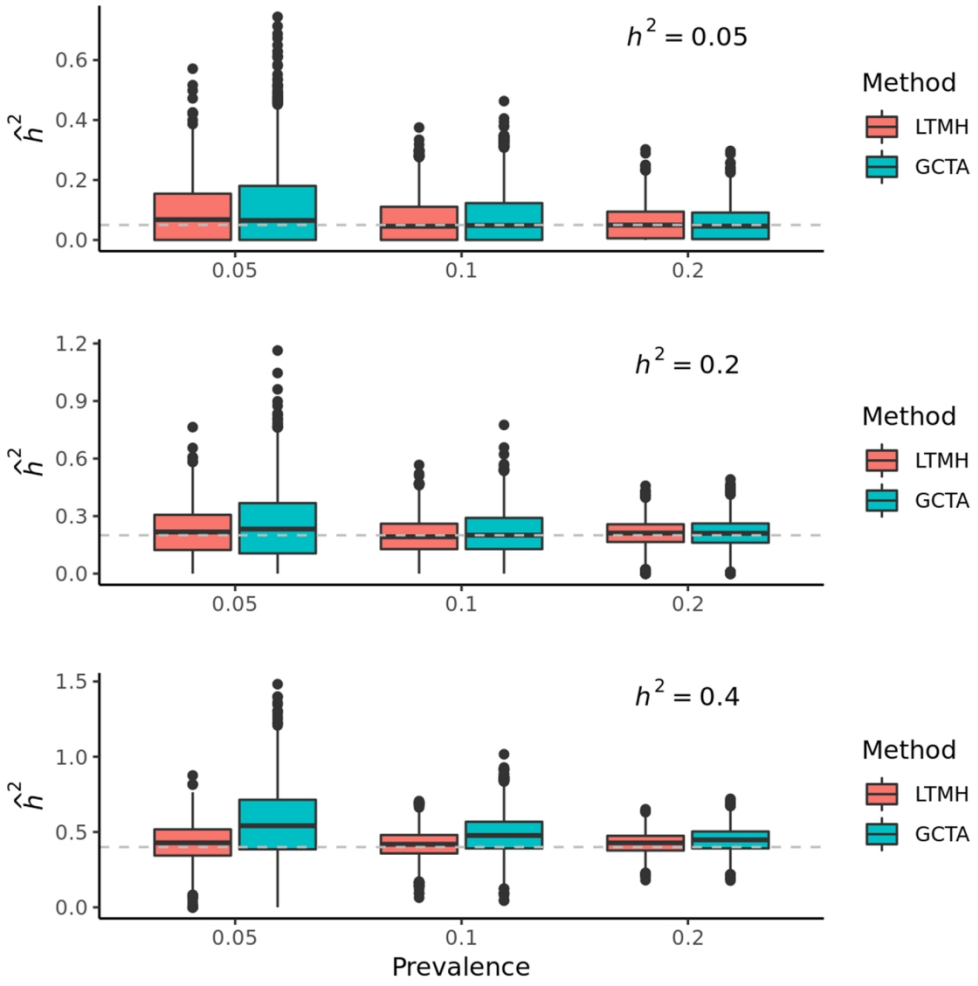

Figure 2. Boxplots for h^2 for randomly selected families (scenario 1). True heritability was 0.05 (top), 0.2 (middle), and 0.4 (bottom) and was indicated as a gray dashed line

152x152mm (300 x 300 DPI)

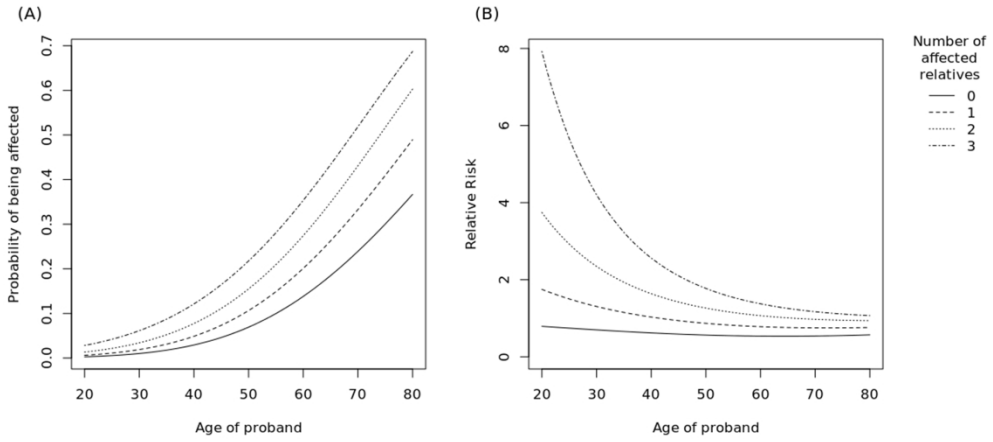

Figure 3. Estimation of risks for T2D according to age. For a certain individual, we assume that he/she has two parents and one younger sibling, and the risk of T2D development was calculated as a function of his/her age and the number of affected family members. The X-axis indicates age of individual, and the age of his/her father and mother were assumed to be 29 years old. The younger sibling was assumed to be 3 years younger than the participant. h^2 and the coefficient of unstandardized age were set to be 0.2944 and 0.051, respectively. (A) Probability of the participant being affected according to the number of affected family members, and (B) relative risks of being affected according to the number of affected family members.

304x147mm (256 x 256 DPI)