Estimating disease liability from age, sex and family history

Jimmy Z Liu (jliu@nygenome.org)

June 20, 2016

1 Introduction

Using a an individual's disease liability as a quantitative phenotype of interest rather than case-control status may boost power to discover risk loci in genetic association studies. Here, we describe how to estimate disease liability given family history of disease and known estimates of disease heritability and prevalence across age and sex.

Conceptually, using demographic and family history information to construct a liability score allows for up-weighting individuals who may have a greater burden of risk-increasing variants and down-weighting those with a lower burden. For instance, an unaffected individual who is old and has no affected relatives will have a lower disease liability than an unaffected individual who is young but whose parents are affected.

2 Methods

2.1 Liability threshold model

An individual's disease liability, l, is a latent continuous variable made up of the sum of environmental and additive genetic components for disease risk. When an individual's liability exceeds a certain threshold, they are said to be affected with that disease. In a population, $l \sim N(0,1)$. The threshold above which individuals are affected with disease is defined as $T = F^{-1}(1-K)$ where F^{-1} is the inverse CDF of the standard normal distribution and K is the disease prevalence. The mean liability of affected individuals is $i = z/K$ and unaffected individuals is $-z(1-K) = -iK(1-K)$, where z is the height of standard normal pdf at threshold T (ref Lee AJHG).

The distribution of individuals who have affected relatives will be shifted to the right compared with the population. The size of this shift depends on h^2, the heritability of disease liability. For instance, the distribution of individuals who have one affected first degree relative is given by:

$$l_1 \sim N\left(\frac{h^2 i}{2}, \frac{1 - h^4 i(i-T)}{4}\right), \tag{1}$$

with a new truncation point and prevalence:

$$T_1 = \frac{T - h^2 i/2}{\sqrt{1 - h^4 i(i-T)/4}} \tag{2}$$

$$K_1 = 1 - F(T_1). \tag{3}$$

As before, the mean liability of individuals with an affected first-degree relative who are affected themselves is z_1/K_1, and unaffected is $-z_1(1-K_1)$, where z_1 is the height of the pdf of the distribution in equation 1 at T_1.

2.2 Truncated multivariate normal distribution

The above example of calculating a mean liability for groups of cases and controls considers only a single affected first relative. We now extend this approach to any combination of affected and unaffected relatives across any pedigree structure. We do this by sampling from the truncated multivariate normal distribution (TMVN).

For a given pedigree with n individuals, let $\boldsymbol{l} = [l_1, l_2...]$ and $\boldsymbol{k} = [k_1, k_2...]$ be the length-n vectors of liabilities and disease prevalences respectively. We vary the values of k_i so that it reflects the disease prevalence across different categories of age and sex (or indeed any number of categories) for individual i. The distribution of $\boldsymbol{l}$ is:

$$\boldsymbol{l} \sim TMVN(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{a}, \boldsymbol{b}), \tag{4}$$

with the pdf given by:

$$f(\boldsymbol{l}, \boldsymbol{\Sigma}, \boldsymbol{a}, \boldsymbol{b}) = \frac{exp\{\frac{1}{2}\boldsymbol{l}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{l}\}}{\int_{\boldsymbol{a}}^{\boldsymbol{b}} exp\{\frac{1}{2}\boldsymbol{l}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{l}\} d\boldsymbol{l}}. \tag{5}$$

Here, $\boldsymbol{\Sigma}$ is the $n \times n$ covariance matrix with diagonal elements equal to 1 and off-diagonal $\{i, j\}$ elements equal to $h^2 \times g_{ij}$, where g_{ij} is the genetic relatedness of individuals i and j in the pedigree. For example, in a family quad (two parents: F, M, and two children: C1, C2):

$$\boldsymbol{\Sigma} = \begin{array}{c} \begin{array}{cccc} F & M & C1 & C2 \end{array} \\ \begin{pmatrix} 1 & 0 & 0.5h^2 & 0.5h^2 \\ 0 & 1 & 0.5h^2 & 0.5h^2 \\ 0.5h^2 & 0.5h^2 & 1 & 0.5h^2 \\ 0.5h^2 & 0.5h^2 & 0.5h^2 & 1 \end{pmatrix} \end{array} \begin{array}{c} F \\ M \\ C1 \\ C2 \end{array}.$$

The length-n vectors $\boldsymbol{a} = [a_1, a_2, ...]$ and $\boldsymbol{b} = [b_1, b_2, ...]$ correspond to the lower and upper bounds respectively of the truncated multivariate normal distribution. Let $\boldsymbol{t} = [t_1, t_2...]$ be the length-n vector of truncation points, where $t_i = F^{-1}(1 - k_i)$. When individual i is unaffected with disease, $a_i = -\infty$ and $b_i = t_i$. When individual i is affected with disease, $a_i = t_i$ and $b_i = \infty$.

To estimate a mean liability for an individual given their case-control status, the case-control status of their family members, disease prevalence and heritability, we first generate random samples from equation 5 using the Gibbs sampler. Briefly, for each iteration, the Gibbs sampler samples from conditional univariate distributions $f(l_i|\boldsymbol{l}_{-i}) = f(l_i|l_1, ..., l_{i-1}, l_{i+1}, ..., l_n)$ which are themselves truncated univariate normal distributions.

The mean of the samples is then taken as that individual's liability - which we refer to as their posterior mean liability (PML).

2.3 Implementation

We use the `tmvtnorm` package in `R` to perform the Gibbs sampling. For each individual, the mean from 1000 random samples (with an additional burn-in of 100 samples) is taken to be the PML. The input files consist of a standard .ped format file with an additional column indicating age, as well as a file specifying prevalence rates for males and females at different age groups. For each family, the `kinship2` package is used to process the pedigree and construct the covariance matrix $\mathbf{\Sigma}$. A heritability of disease liability also needs to be defined.

3 Results

We estimated a PML for 116,000 individuals along with their parents and siblings in the UK Biobank across 12 phenotypes. All individuals reported the phenotypes of their mother, father and any siblings. While individuals reported their number of siblings, no information was provided as to which particular sibling was affected by disease, so we just assumed a single single affected or unaffected sibling depending on the reported phenotype. The age of parents were reported as either current age or age at death. No age was reported for siblings, so we approximate the age of the sibling to be the same as the genotyped individual.

For each of the 12 common diseases, prevalences across age and sex were taken directly from the full UK Biobank data of 500,000 individuals. We used both the age of sex of these individuals along with their parents to estimate disease prevalences. Due the small numbers, the ages considered were 40-100. The disease prevalences across the 12 phenotypes are shown in figure 1. The distribution of estimated PMLs are shown in figure 2.

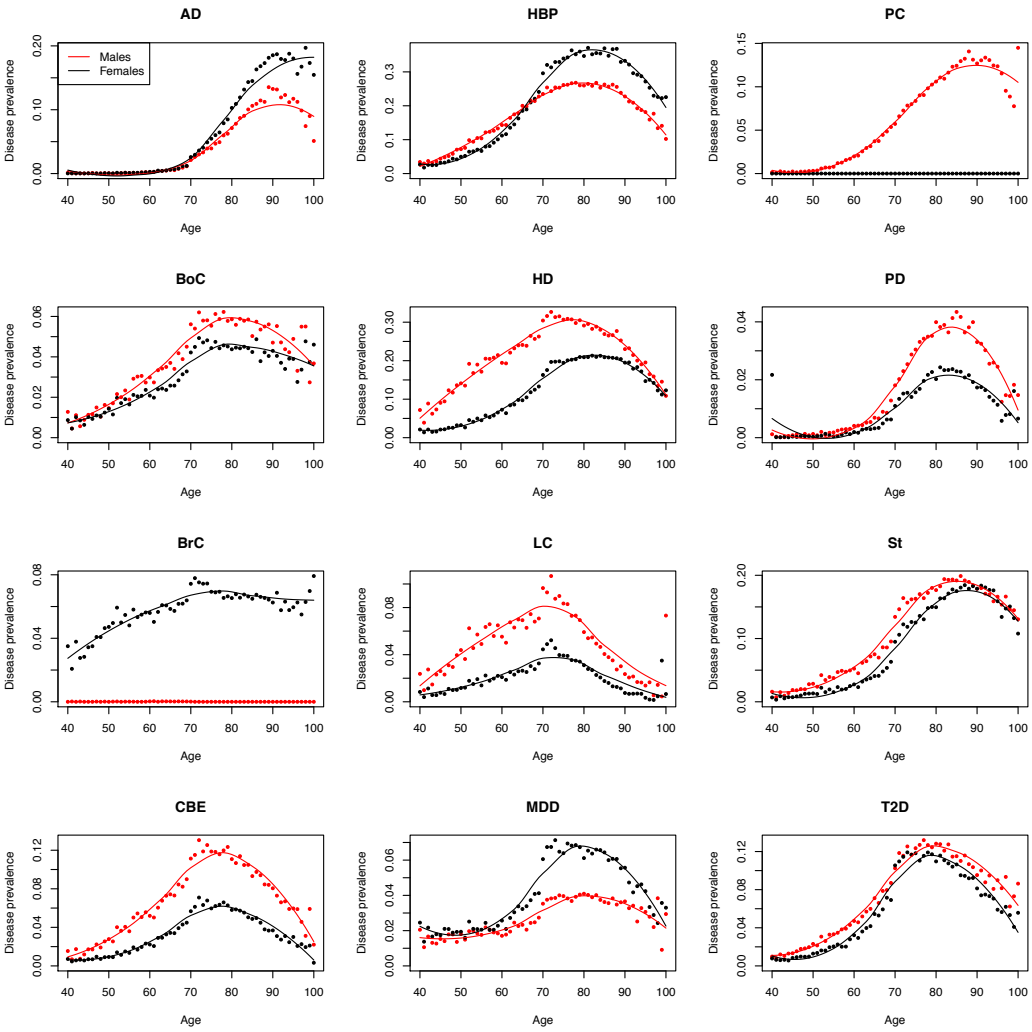

Figure 1. Disease prevalence across age and sex for 12 diseases

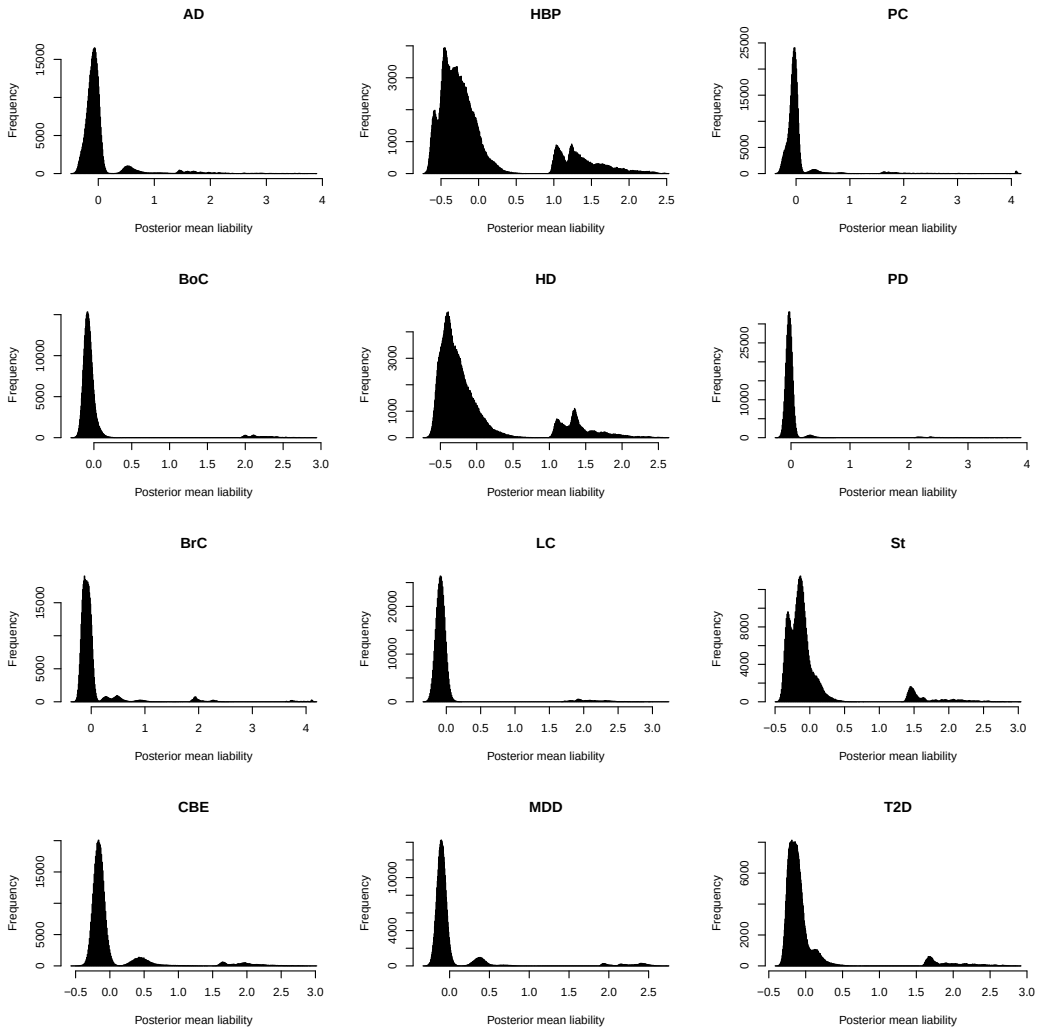

Figure 2. Histogram of posterior mean liabilities across 12 phenotypes