Statistical methods for predicting disease risk from genetics data in the biobank era

Frederick Boehm

Key Points

Abstract

Introduction

Our genes influence many aspects of our health and biology. While all humans share more than 99% of their DNA, every human's DNA contains millions of differences that make it unique. These differences, also called genetic variants, not only give us varying heights and eye colors, but they also influence our risk of developing diseases like cancer, diabetes, and heart disease. Recent technologies have made it possible to acquire genetic data for millions of people, and this data has been used to identify thousands of genetic variants that influence disease risks.

- Multiple ancestries
- Statistical methods
- Incorporating other risk factors: past medical history, family history
- Common vs low frequency alleles

Genome-wide association studies probe genetic markers, one at a time, across the entire genome to identify gene regions where the marker genotypes correlate with disease status. Due to correlation between marker genotypes on a single chromosome, a phenomenon called linkage disequilibrium, it is typical for a group of consecutive markers to all correlate with disease status.

Sharing of genome-wide association study results, typically termed summary statistics, has propeled the field of human genetics forward over the last ten years. It is now common for scientists to freely and publicly share summary statistics from genome-wide association studies. Other teams of investigators may freely use the summary statistics in their own investigations.

In the most general form, a PRS is calculated as the weighted sum of minor allele counts at a collection of genetic markers. PRS methods differ in how they determine the weights and

in which genetic markers are used in the sum. Clumping and thresholding, the earliest PRS method, uses LD patterns among genetic variants to subset the genome-wide collection of markers. The next step in clumping and thresholding selects only those markers with strong disease associations to include in the weighted sum that determines the PRS. Weights for the minor allele counts for the final set of markers often leverage estimated effects from GWAS.

LDpred2

DBSLMM, which abbreviates Deterministic Bayesian sparse linear mixed model, assumes that SNP effects from across the genome arise from a mixture of two normal distributions that differ only in variance. The component with the larger variance accounts for the SNPs with large effects, while the component with the smaller variance accommodates the majority of the genome-wide SNPs, which are assumed to have small effects. DBSLMM then assumes that the SNPs with effects belonging to the large variance distribution can be identified by standard GWAS methods, since they tend to have large effects. The remaining SNPs are modeled jointly with the small variance distribution and treated as a "polygenic" effect. While we have little statistical power to accurately estimate the effects for the individual SNPs in the small variance distribution, DBSLMM demonstrates the ability to collectively model the polygenic effect from all of the small effect SNPs with modest accuracy. This accounts for the better predictive performance of DBSLMM compared to, for example, clumping and thresholding.

PRSCS

Open questions in PRS methods

Despite great advances in PRS methods over the last decade, a number of important questions remain. Many of these questions must be answered before we can truly derive the greatest health benefits from genetics data. These open questions include:

1. Given the poor performance of PRS for subjects with non-European ancestry, how do we calculate PRS for any subject, regardless of ancestry?
2. With the clinical successes of disease risk predictors like the Framingham survey for cardiovascular event prediction, how do we synergize clinical risk assessment with genetic risk assessment to produce even more accurate predictions?
3. With most PRS methods modeling only additive genetic effects, how do we incorporate gene-gene and gene-environment interactions to achieve better predictions?
4. With these PRS methods all being new, and much of the public being skeptical of genetic technologies, how do we effectively communicate the benefits of PRS risk predictions to patients and clinicians?

References