Predicting disease risk from genetics data in the biobank era

Key Points

1. Polygenic risk scores hold promise as future clinical tools to quantify disease risk.
2. A polygenic risk score is disease-specific and tries to summarize in a single number a person's genetic susceptibility to a disease.
3. While there are many statistical methods for constructing polygenic risk scores, they all use weighted sums of minor allele counts at genetic variants.
4. Advances in technology have enabled affordable acquisition of genetic and phenotypic data from millions of people.
5. The large volume of available data for polygenic risk scores requires new statistical methods with efficient computational implementations.

Abstract

I present a brief review of polygenic risk scores in the biobank era. A polygenic risk score for a specified disease aims to summarize and quantify a person's genetic predisposition to the disease. Advances in technology have enabled affordable acquisition of genetic and phenotypic data from millions of people across the globe. Over the last ten years, widespread data sharing and an ever-growing volume of genetic data have motivated computationally scalable and efficient methods for polygenic risk score construction. Despite their great promise, scientists have yet to fully develop polygenic risk scores for use in clinical medicine and public health. I conclude the review with a discussion of challenges to surmount in order to advance this line of research.

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.

More recently, the human genetics research community has invested tremendous time and resources in strategies and tools to quantify the disease risk for individual persons. If we know the genetic variants in a person's DNA, can we predict their risk for, for example, coronary artery disease? The most fruitful line of investigation has summarized genetic variants' effects into a single number, termed a polygenic risk score, or PRS, for each person. Because the polygenic risk score for a specified disease is derived from the genetic variants present in nearly every human cell in a person, the polygenic risk score can be calculated sometimes decades before disease onset. Thus, polygenic risk scores offer the possibility of identifying people who are genetically predisposed towards developing a disease. These people may benefit from additional interventions, such as earlier or more frequent disease screening. At the same time, polygenic scores alone don't foretell a future disease; rather, people with higher polygenic risk scores tend to have greater risk of eventually developing the disease.

A rapid increase in genetic data availability has enabled construction and testing of polygenic risk scores for hundreds of diseases. Because most genome-wide association studies (GWAS) use genetic data from single nucleotide polymorphism (SNP) arrays, PRS tend to use SNP genotypes while not explicitly using other classes of genetic variants. Typically, SNP effect estimation and polygenic risk score construction are performed in one set of subjects, the training set, while polygenic risk score assessment involves a separate, non-overlapping, set of subjects, the test set, to reduce statistical biases.

Increasing data availability & growth of biobanks

Development of tissue repositories, also called biobanks, with samples from tens or hundreds of thousands of people has accelerated the demand for computationally efficient statistical methods for PRS construction (Collister, Liu, and Clifton 2022). The most widely used biobank to date, the UK Biobank, contains tissue samples for nearly 500,000 British adults (Bycroft et al. 2018). Investigators around the world apply for access to UK Biobank data for their research.

Similar biobank efforts are underway in Finland (FinnGen) (Kurki et al. 2023), Estonia (Leitsalu et al. 2015), Japan (Nagai et al. 2017), the USA (All of Us (Us Research Program Investigators 2019) and Million Veterans Program (Gaziano et al. 2016)), and other locales around the world. They all curate genetic and phenotypic data for their study subjects and share it for research purposes.

The different biobanks offer, at a minimum, genetic and phenotypic data for tens of thousands of people, with some, such as the UK Biobank, containing data for hundreds of thousands. Available phenotypes also differ among biobanks. While the UK Biobank measures tens of thousands of traits - including, for some people gene expression levels and protein abundances, among other molecular traits - some biobanks have measured other traits. Fortunately, many biobanks are using tissue samples to acquire new traits in response to advances in biotechnology.

Sharing GWAS summary statistics

In the last ten years, many investigators have shared summary statistics from GWAS Buniello et al. (2019). While it is often logistically difficult to share individual-level genetic and phenotypic data, sharing of GWAS summary statistics, incuding the genome-wide SNP effect estimates and their variances and sometimes the linkage disequilibrium patterns for the study subjects, requires less memory storage and avoids many data privacy concerns relative to sharing of individual-level data (MacArthur et al. 2021).

Some research funders adopted a requirement that investigators publicly share GWAS summary statistics (Thelwall et al. 2020). This requirement has promoted widespread availability of GWAS summary statitics for thousands of traits in diverse study populations. One particularly useful resource is the GWAS summary statistics for thousands of UK Biobank traits from a research team at the Broad Institute. This development also has motivated PRS methods that use GWAS summary statistics and don't use individual-level data Mak et al. (2017).

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.

Methods for polygenic risk score construction

In the most general form, a PRS is calculated as the weighted sum of minor allele counts at a collection of genetic markers (Ma and Zhou 2021). PRS methods differ in how they determine the weights and in which genetic markers are used in the sum. Three popular methods are clumping and thresholding Privé et al. (2019), LDpred2 (Privé, Arbel, and Vilhjálmsson 2020), and DBSLMM (S. Yang and Zhou 2020). 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.

While clumping and thresholding is easy to implement and computationally fast, researchers have found prediction performance gains with models that impose additional assumptions on SNP effect distributions. For example, LDpred2 depends on a Bayesian model for SNP effects. It assumes that a proportion of the SNPs affect the disease status, while many have no effect. Those SNPs that impact disease status have effects that follow a normal distribution. The LDped2 developers use a probabilistic sampling approach called Gibbs sampling for model fitting and SNP effect estimation. Ease of use and computing speed, as well as predictive performance, have contributed to widespread use of 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.

While many other PRS construction methods exist, I've limited discussion here to three of the most popular strategies. Ma and Zhou (2021), writing in 2021, offer a deeper review of construction methods.

Pipeline for polygenic risk score construction and applications

Ma and Zhou (2021), in their review of polygenic risk score methods, presented a flow diagram to summarize the construction and application of polygenic risk scores Figure 1. They divide the procedures into six sequential steps: data input, data processing, model fitting, validation, testing, and applications. The data input step involves identification of a suitable training set of subjects. Often, privacy concerns force investigators to share only GWAS summary statistics, such as the effect estimates and their variances, so it is important that the chosen construction method accept summary statistics as inputs. A test set of subjects, for whom individual-level data is typically available, are also selected in this step. Data processing follows, where investigators perform quality control procedures on the training and test set data (Privé et al. 2022). In the model fitting step, researchers use a PRS construction method, like clumping and thresholding or DBSLMM, to estimate SNP effect sizes from the training data. Because some PRS construction methods, such a LDpred2, use models with statistical hyperparameters, Ma and Zhou (2021) insert an optional validation step for tuning model hyperparameters as needed. The testing step uses the test set data to evaluate the PRS performance. Using the SNP effect size estimates from the model fitting step, researchers calculate a polygenic risk score for each test set subject and use them to evaluate predictive performance through statistics such as area under the curve and pseudo-R^2. Following testing, many researchers apply the newly developed polygenic risk scores. Risk stratification orders the subjects by polygenic risk score, then evaluates potential interventions for those people with the highest polygenic risk scores. The exact threshold for categorizing a person as "high-risk" for a specified disease varies with disease and application goals, but often the top ten percent receives the "high-risk" designatin. Importantly, polygenic risk scores do not lie only between 0 and 1, and, thus, do not directly correspond to a subject's absolute risk of disease.

Current limitations of polygenic risk scores

Clinical applications of polygenic risk scores have been limited (Torkamani, Wineinger, and Topol 2018). Because most GWAS use European ancestry subjects, most polygenic risk scores use European ancestry subjects for SNP effect estimation (Buniello et al. 2019). However, most people don't have European ancestry. In fact, scientists have observed that polygenic risk scores built from data from subjects with European ancestry often don't perform well in people with non-European ancestry (Dikilitas et al. 2020). While the sources of these performance discrepancies remains a subject for research, many scientists partially attribute it to differing patterns in linkage disequilibrium among ancestral groups (Slatkin 2008). For example, those with (recent) African ancestry tend to have weaker linkage disequilibrium on each chromosome compared to those with European ancestry. Wang et al. (2022) discussed the issues of equity and fairness in the context of clinical application of polygenic risk scores.

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 (Duncan et al. 2019)?
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 (Mahmood et al. 2014)?
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 (Palk et al. 2019)?

References

Barbeira, Alvaro N, Scott P Dickinson, Rodrigo Bonazzola, Jiamao Zheng, Heather E Wheeler, Jason M Torres, Eric S Torstenson, et al. 2018. "Exploring the Phenotypic Consequences of Tissue Specific Gene Expression Variation Inferred from GWAS Summary Statistics." *Nature Communications* 9 (1): 1825.

Buniello, Annalisa, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." *Nucleic Acids Research* 47 (D1): D1005–12.

Bycroft, Clare, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, et al. 2018. "The UK Biobank Resource with Deep Phenotyping and Genomic Data." *Nature* 562 (7726): 203–9.

Collister, Jennifer A, Xiaonan Liu, and Lei Clifton. 2022. "Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists." *Frontiers in Genetics* 13: 818574.

Dikilitas, Ozan, Daniel J Schaid, Matthew L Kosel, Robert J Carroll, Christopher G Chute, Joshua C Denny, Alex Fedotov, et al. 2020. "Predictive Utility of Polygenic Risk Scores for Coronary Heart Disease in Three Major Racial and Ethnic Groups." *The American Journal of Human Genetics* 106 (5): 707–16.

Duncan, Laramie, H Shen, B Gelaye, J Meijsen, K Ressler, M Feldman, R Peterson, and Ben Domingue. 2019. "Analysis of Polygenic Risk Score Usage and Performance in Diverse Human Populations." *Nature Communications* 10 (1): 3328.

Gaziano, John Michael, John Concato, Mary Brophy, Louis Fiore, Saiju Pyarajan, James Breeling, Stacey Whitbourne, et al. 2016. "Million Veteran Program: A Mega-Biobank to Study Genetic Influences on Health and Disease." *Journal of Clinical Epidemiology* 70: 214–23.

Kurki, Mitja I, Juha Karjalainen, Priit Palta, Timo P Sipilä, Kati Kristiansson, Kati M Donner, Mary P Reeve, et al. 2023. "FinnGen Provides Genetic Insights from a Well-Phenotyped Isolated Population." *Nature* 613 (7944): 508–18.

Leitsalu, Liis, Toomas Haller, Tõnu Esko, Mari-Liis Tammesoo, Helene Alavere, Harold Snieder, Markus Perola, et al. 2015. "Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu." *International Journal of Epidemiology* 44 (4): 1137–47.

Lloyd-Jones, Luke R, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E Kemper, Huanwei Wang, et al. 2019. "Improved Polygenic Prediction by Bayesian Multiple Regression on Summary Statistics." *Nature Communications* 10 (1): 5086.

Ma, Ying, and Xiang Zhou. 2021. "Genetic Prediction of Complex Traits with Polygenic Scores: A Statistical Review." *Trends in Genetics* 37 (11): 995–1011.

MacArthur, Jacqueline AL, Annalisa Buniello, Laura W Harris, James Hayhurst, Aoife McMahon, Elliot Sollis, Maria Cerezo, et al. 2021. "Workshop Proceedings: GWAS Summary Statistics Standards and Sharing." *Cell Genomics* 1 (1).

Mahmood, Syed S, Daniel Levy, Ramachandran S Vasan, and Thomas J Wang. 2014. "The

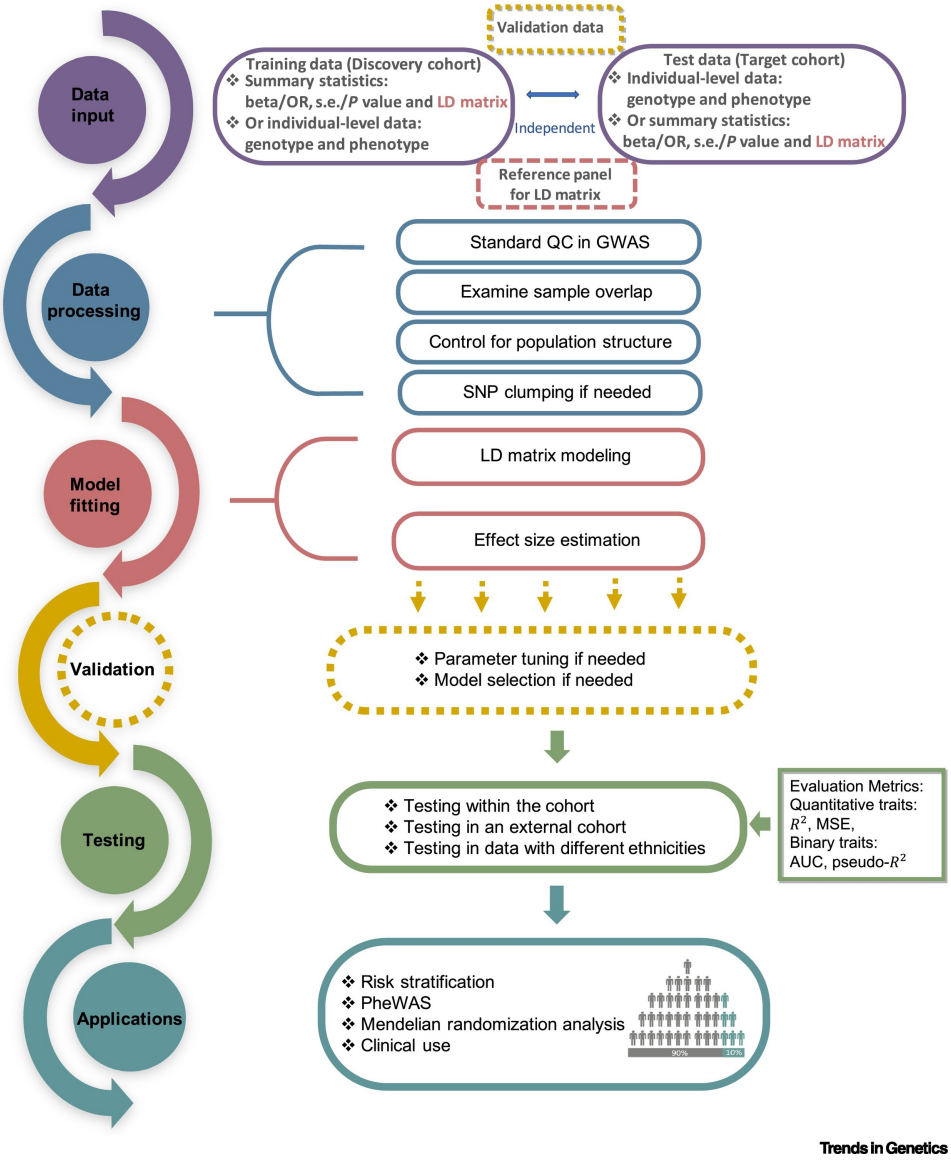

Figure 1: Flow diagram for PRS analysis from Ma & Zhou (2021).

Framingham Heart Study and the Epidemiology of Cardiovascular Disease: A Historical Perspective." *The Lancet* 383 (9921): 999–1008.

Mak, Timothy Shin Heng, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. 2017. "Polygenic Scores via Penalized Regression on Summary Statistics." *Genetic Epidemiology* 41 (6): 469–80.

Nagai, Akiko, Makoto Hirata, Yoichiro Kamatani, Kaori Muto, Koichi Matsuda, Yutaka Kiyohara, Toshiharu Ninomiya, et al. 2017. "Overview of the BioBank Japan Project: Study Design and Profile." *Journal of Epidemiology* 27 (Supplement_III): S2–8.

Palk, Andrea C, Shareefa Dalvie, Jantina De Vries, Alicia R Martin, and Dan J Stein. 2019. "Potential Use of Clinical Polygenic Risk Scores in Psychiatry–Ethical Implications and Communicating High Polygenic Risk." *Philosophy, Ethics, and Humanities in Medicine* 14 (1): 1–12.

Privé, Florian, Julyan Arbel, Hugues Aschard, and Bjarni J Vilhjálmsson. 2022. "Identifying and Correcting for Misspecifications in GWAS Summary Statistics and Polygenic Scores." *Human Genetics and Genomics Advances* 3 (4).

Privé, Florian, Julyan Arbel, and Bjarni J Vilhjálmsson. 2020. "LDpred2: Better, Faster, Stronger." *Bioinformatics* 36 (22-23): 5424–31.

Privé, Florian, Bjarni J Vilhjálmsson, Hugues Aschard, and Michael GB Blum. 2019. "Making the Most of Clumping and Thresholding for Polygenic Scores." *The American Journal of Human Genetics* 105 (6): 1213–21.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3): 559–75.

Slatkin, Montgomery. 2008. "Linkage Disequilibrium—Understanding the Evolutionary Past and Mapping the Medical Future." *Nature Reviews Genetics* 9 (6): 477–85.

Thelwall, Mike, Marcus Munafò, Amalia Mas-Bleda, Emma Stuart, Meiko Makita, Verena Weigert, Chris Keene, Nushrat Khan, Katie Drax, and Kayvan Kousha. 2020. "Is Useful Research Data Usually Shared? An Investigation of Genome-Wide Association Study Summary Statistics." *Plos One* 15 (2): e0229578.

Torkamani, Ali, Nathan E Wineinger, and Eric J Topol. 2018. "The Personal and Clinical Utility of Polygenic Risk Scores." *Nature Reviews Genetics* 19 (9): 581–90.

Us Research Program Investigators, All of. 2019. "The 'All of Us' Research Program." *New England Journal of Medicine* 381 (7): 668–76.

Wang, Ying, Kristin Tsuo, Masahiro Kanai, Benjamin M Neale, and Alicia R Martin. 2022. "Challenges and Opportunities for Developing More Generalizable Polygenic Risk Scores." *Annual Review of Biomedical Data Science* 5: 293–320.

Yang, Jian, Teresa Ferreira, Andrew P Morris, Sarah E Medland, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Pamela AF Madden, et al. 2012. "Conditional and Joint Multiple-SNP Analysis of GWAS Summary Statistics Identifies Additional Variants Influencing Complex Traits." *Nature Genetics* 44 (4): 369–75.

Yang, Sheng, and Xiang Zhou. 2020. "Accurate and Scalable Construction of Polygenic Scores

in Large Biobank Data Sets." *The American Journal of Human Genetics* 106 (5): 679–93.