Specific Aims

Polygenic risk scores use genetic data to quantify disease risk and thus provide clinicians and public health workers with additional risk factors to inform preventive efforts with the goals of enhancing health, lengthening life, and reducing illness and disability. Recent technology advances have enabled genotyping and phenotyping of hundreds of thousands of individuals as biomedical scientists seek to identify the genetic underpinnings of complex diseases in studies called genome-wide association studies. Genome-wide association studies estimate associations between SNPs, one at a time, and a complex disease. Polygenic risk scores leverage genome-wide SNP effects from genome-wide association studies to quantify disease risk. While scientists have developed many statistical methods to calculate polygenic risk scores, current methods ignore information that may improve predictive performance. Currently, a major obstacle in the field is that polygenic risk scores neglect gene-gene interactions.

To solve this problem, we will develop a polygenic risk score method that accounts for gene-gene interactions in a Bayesian statistical model. In our Bayesian model, we will specify a spike-and-slab prior distribution for genome-wide SNP effects and SNP-SNP interactions. A spike-and-slab prior distribution assumes that most effects are zero, while a small proportion of effects arise from a normal distribution. We'll use mean field variational inference methods to approximate the posterior distribution in an accurate and computationally efficient manner. The feasibility of our strategy is supported by recent advances in variational inference and in large-scale computing. Our proposed statistical method is expected to outperform current polygenic risk score methods that ignore gene-gene interactions. Our new, more accurate method for calculating polygenic risk scores will enable better stratification of individuals by genetic risk, which, in turn, will allow for more accurate targeting of preventive public health interventions.

Aim 1 will develop a Bayesian statistical model for polygenic risk scores that uses only SNP effect estimates (and ignores SNP-SNP interactions). We'll use a spike-and-slab prior for SNP effects across the genome to induce sparsity. We'll use mean field variational methods to fit the posterior distribution for SNP effects. With the posterior estimates of SNP effects, we'll construct polygenic risk scores as weighted sums of SNP minor allele counts, where the weights are the posterior SNP effect estimates.

Aim 2 will develop a Bayesian statistical model for polygenic risk scores based on SNP effect estimates and estimates for SNP-SNP interactions. We'll use a spike-and-slab prior to induce sparsity of both SNP effects and SNP-SNP interactions. We'll fit our Bayesian model by using approximations from mean field variational methods. We'll write open-source computer code to implement our method, then assess its statistical performance with simulated traits.

Our proposed studies will establish mean field variational methods as a computationally efficient and accurate method for posterior inference in biobank-scale genetics studies, an approach that current methods lack. Furthermore, our new methods will enable the modeling of gene-gene interactions to improve prediction accuracy. Our open-source, well documented, and thoroughly tested software will provide the genetics research community with a valuable resource and enable future advances in statistical methods for polygenic risk score construction. Our studies will improve accuracy of ongoing public health efforts to identify people at high risk for deadly diseases and will enable early, preventive interventions to improve the lives of millions around the world.

Significance

Importance of the Problem to be Addressed

While there are many statistical methods that calculate polygenic risk scores from GWAS summary statistics, current approaches have limited predictive ability. For example, among psychiatric conditions, polygenic risk scores predict only 2% of the liability variance for major depressive disorder (Wray et al. 2018), 5% for bipolar disorder (Mullins et al. 2021), 3% for neuroticism (Luciano et al. 2018), and 6% for attention deficit hyperactivity disorder (Demontis et al. 2019). The statistical methods that underlie these polygenic risk score calculations all share the assumption that each SNP has only a genetically "additive" effect on the trait. In other words, the methods assume that the trait liability is a weighted sum of minor allele counts at a collection of SNPs, with weights specified by the GWAS estimate of the SNP effect. By making the simplifying assumption that SNP-SNP interactions have no net impact on the trait liability, the investigators restrict the predictive ability of the polygenic risk scores. While the incorporation of SNP-SNP interactions into polygenic risk score calculations doesn't fully resolve the limited predictive ability of polygenic risk scores, including the possibility of SNP-SNP interactions in polygenic risk score calculations will improve predictive ability over methods that neglect SNP-SNP interactions. The reason for this is that the collection of statistical models with possible SNP-SNP interactions contains the collection of statistical models without SNP-SNP interactions.

By improving predictive ability of polygenic risk scores beyond current standards, investigators will more accurately target interventions and preventive measures to those individuals at highest risk of disease. Clinical researchers who use polygenic risk scores to counsel patients at high risk for disease will more accurately identify high-risk patients, while public health researchers will more accurately identify high-risk populations for targeted preventive measures. Together, these efforts will positively impact society by enhancing health, lengthening life, and reducing illness and disability.

Failure to address this issue, by continuing to use current polygenic risk score methods that ignore contributions from SNP-SNP interactions, will result in misclassification of individuals into high-risk and low-risk categories. Misclassification of individuals will attenuate measures of intervention effectiveness, since many of the subjects classified as "high-risk", in fact, will truly be low-risk. Thus, interventions that reduce disease burden and extend healthy lifespan in populations will not be widely implemented and potential gains in lifespan and well-being will not be achieved.

Rigor of the Prior Research Supporting the Aims

Aim 1 (Literature)

Technology advances have fueled a resurgence in interest in Bayesian statistical methods. While historically Bayesian approaches required time-consuming, computationally intense sampling methods for model fitting and inference, high-performance computers and analytic approximation methods for Bayesian models have accelerated the widespread adoption of Bayesian models for high-dimensional data(Blei, Kucukelbir, and McAuliffe 2017). Development of variational inference methods for Bayesian modeling has been especially fruitful. They have propeled analysis of extremely large data sets, including those where traditional linear regression models fail, as when the number of variables is larger than the sample size.

Zhang, Xu, and Zhang (2019) developed a variational inference strategy for sparse model selection in logistic regression. In other words, their work imposes the modeling assumption that most SNPs have no effect on the trait, while the remaining small proportion of SNPs all impact the trait. Ray, Szabo, and Clara (2020) establish a theoretical foundation for our proposal to use a Bayesian model for high-dimensional logistic regression. In our setting, disease status is the binary outcome variable, while the millions of SNP genotypes per subject are the high-dimensional, dependent

variables in the logistic regression model. Ray, Szabo, and Clara (2020) provide an efficient mean field variational inference approximation that enables accurate variable selection and model inferences.

Aim 2 (Literature)

The imposition of sparsity assumptions in our statistical modeling is especially relevant when we incorporate SNP-SNP interactions into our polygenic risk score modeling. This is due to the large number of possible SNP-SNP interactions across the genome. Even if we make the reasonable assumption that SNP-SNP interactions for SNPs on distinct chromosomes have no effect on the trait, consideration of interactions involving only Chromosome 1 SNPs exceeds 10^{10}, due to the large number of SNPs on Chromosome 1. Thus, we need to impose a sparsity-inducing assumption. In other words, we choose to assume that the majority of SNP-SNP interactions have no effect on the trait. We then let our Bayesian model "learn" which SNP-SNP interactions have non-negligible trait effects.

Significance of the Expected Research Contribution

Upon successful completion of the proposed research, we expect our contribution to be a computationally efficient and scalable new statistical method for calculating accurate polygenic risk scores. *This contribution is expected to be significant because it will enable accurate identification of individuals at high disease risk and appropriate targeting of preventive interventions.* We will freely share our open source, well documented, and thoroughly tested software implementation of our statistical methods to benefit the research community and those who want to build upon our findings. Society will benefit from our research through the clinical and public health preventive measures that our work makes possible. Ultimately, our research will promote health, lengthen life, and reduce illness and disability.

Innovation

The status quo as it pertains to polygenic risk score construction is to neglect contributions of SNP-SNP interactions to trait values. We depart from the status quo by allowing for and modeling SNP-SNP interactions in construction of polygenic risk scores. The modest predictive ability of current polygenic risk score methods motivates the need for advances like those that we propose. Without improvements in polygenic risk score construction, proposed interventions that target those with high polygenic risk scores for a disease of interest will inaccurately label many individuals as high risk, and thus misallocate clinical and public health resources.

Our strategy to include SNP-SNP interactions is supported by previous reports that, for some diseases, SNP-SNP interactions contribute considerably to the trait values (Li et al. 2015). The contributions of epistasis to trait values and disease risks motivates the modeling of SNP-SNP interactions. Bateson (1909) coined the term "epistasis" and Fisher (1919) developed its modern definition, where epistasis is the departure of a trait value from that expected under the "additive" genetic model. Fisher (1919) also recognized the role of epistasis on quantitative trait values. Lee et al. (2020) shared an algorithm to detect SNP-SNP interactions in GWAS data sets and applied it to schizophrenia study to identify SNP-SNP interactions in biologically plausible gene pathways.

The proposed research is innovative, in our opinion, because it represents a substantive departure from the status quo by developing and assessing polygenic risk score methods that leverage SNP-SNP interactions in addition to SNP main effects.

Our proposed research opens new horizons in clinical, public health, and biostatistical research. In clinical research, the more accurate polygenic risk scores that will result from our work will lead to more accurate identification of individuals at high risk for disease. This, in turn, will allow more accurate measurement of effects of clinical interventions on disease development.

Public health research benefits, like those in clinical research, arise from more accurate classification of individuals as being at high risk for disease. Specifically, the identification of collections of collections of high risk subjects will enable the tailoring of preventive interventions to reduce disease burden among the most vulnerable.

Biostatistics, and specifically the field of statistical genetics, will achieve new horizons through our research because we will be the first to characterize the performance of polygenic risk score methods that use SNP-SNP interactions. The expected improvement in predictive accuracy for our methods means that they will be widely adopted by other researchers. Furthermore, our research products, such as our highly modular, well documented, and thoroughly tested software products will provide a springboard for further innovations in polygenic score construction methods by ourselves and other research teams.

Approach

Overall Research Design

Specific Aim 1: We will use mean field variational methods to provide analytic approximations to the posterior distribution for a Bayesian model with sparsity-inducing priors for polygenic risk scores.

Introduction

Current inference methods for Bayesian polygenic risk score models use sampling-based strategies like Markov chain monte carlo and, thus, pose significant computational burdens for modern human genetics studies with large sample sizes and high-dimensional measurements. The objective of Specific Aim 1 is to use computationally efficient and scalable variational inference methods for posterior inference in polygenic risk score models. Variational inference uses an analytic approximation to the posterior distribution to estimate quantities of interest. To achieve this objective, we will use the polygenic risk score model developed Privé, Arbel, and Vilhjálmsson (2020). However, instead of using the sampling-based inference methods of Privé, Arbel, and Vilhjálmsson (2020), we will apply the variational inference methods of Ray, Szabo, and Clara (2020) and Yang, Pati, and Bhattacharya (2020) for approximate inferences from the posterior distribution. The rationale is that our variational inference-based strategy will diminish computing time and memory requirements while maintaining predictive ability of the sampling-based strategy of Privé, Arbel, and Vilhjálmsson (2020). Moreover, with the need to model polygenic risk scores from studies with sample sizes nearing one million subjects, each of which has thousands of phenotype measurements, current sampling-based strategies for posterior inference are inadequate. Variational methods, on the other hand, are computationally efficient and scalable to big data sets. We will compare our method's performance - in terms of predictive ability and computing resource requirements - against that of Privé, Arbel, and Vilhjálmsson (2020). Upon completion of Specific Aim 1, it is our expectation that we will have created a computationally scalable and efficient method for constructing polygenic risk scores. Our open source implementation ensures transparency in our research and provides a valuable analytic tool to human genetics researchers.

Research Design

Study Data

We will use imputed genotype and phenotype data from the UK Biobank Study (Bycroft et al. 2018). The UK Biobank study enrolled approximately 500,000 UK adults. Each subject has tens of thousands of phenotypic measurements. The UK Biobank Study shares protected individual-level data with investigators around the world through its data sharing agreement. For polygenic risk score construction, we will restrict our genetic markers to those available in the Hapmap3 Study (Consortium et al. 2010), like Privé, Arbel, and Vilhjálmsson (2020) and Ge et al. (2019). This will afford us 1,117,493 across the genome. We will restrict the UK Biobank subject set to those used in principal components analysis, who are unrelated and passed quality control filters (Privé, Arbel, and Vilhjálmsson 2020). This will leave us with 362,320 subjects (Privé, Arbel, and Vilhjálmsson 2020).

Statistical modeling

We will use the Bayesian statistical models in LDpred (Vilhjálmsson et al. 2015) and LDpred2 (Privé, Arbel, and Vilhjálmsson 2020). LDpred assumes that the SNP (main) effects follow a distribution that is a mixture of a normal distribution and a point mass at zero. In mathematical notation,

$$\beta_j \sim \begin{cases} N(0, \frac{h^2}{Mp}), & \text{with probability } p \\ 0, & \text{with probability } 1-p \end{cases}$$

where β_j is the SNP effect for SNP with index j, M is the number of SNPs across the genome, p is the proportion of SNPs that are causal for the disease, and h^2 is the SNP heritability of the disease (Vilhjálmsson et al. 2015).

Statistical inference methods

Because Bayesian statistical models routinely have intractible posterior distributions, researchers are often forced to construct elaborate sampling-based strategies that rely on Markov chain monte carlo or related methods like Gibbs sampling. However, recent advances in the mathematics of variational inference have provided alternatives for posterior inference in mathematically intractible Bayesian models. Mean field variational inference for Bayesian statistical models imposes a simplifying assumption and results in a

Assessing predictive performance

Following procedures of Privé, Arbel, and Vilhjálmsson (2020), we will assign 10,000 subjects to a "validation" set of subjects. We will use the validation set to tune model hyperparameters and to estimate genome-wide SNP-SNP correlations. With the remaining 352,320 subjects, we will randomly assign 300,000 for use in our genome-wide association studies, which are a prerequisite for our polygenic risk score calculations (Privé, Arbel, and Vilhjálmsson 2020). The remaining 52,320 subjects that are in neither the GWAS cohort nor the validation set, are assigned to the "test" set (Privé, Arbel, and Vilhjálmsson 2020). We will use the test set to evaluate the performance of our polygenicscores. To compare our proposed method with those from Privé, Arbel, and Vilhjálmsson (2020) and Ge et al. (2019), we will compute the area under the receiver operating characteristic curve for all methods. The receiver operating characteristic curve plots the performance of a classifier, like our polygenic risk scores used to classify subjects as disease cases or controls, across a range of classification thresholds. The area under the receiver operating characteristic curve is one measure of the method's predictive performance. We will follow the detailed procedure de-

scribed by Privé, Arbel, and Vilhjálmsson (2020) by sampling 10,000 bootstrap replicates of the test set subjects and computing the area under the receiver operating characteristic curve for each bootstrap replicate. With the resulting 10,000 areas, we will report the mean, the 2.5 percentile, and the 97.5 percentile. Privé et al. (2018) have implemented this strategy in the user-friendly R package, `bigstatsr`.

Expected Outcomes, Potential Problems & Alternative Strategies

Our expected outcomes from Specific Aim 1 include a new statistical method for polygenic risk scores. Unlike existing methods, we expect that our method will be computationally efficient and scalable to data sets with millions of subjects and thousands of traits.

One possible problem lies in our use of mean field variational inference instead of other variational inference approaches. Mean field variational inference is equivalent to α-variational inference with $\alpha = 1$ (Yang, Pati, and Bhattacharya 2020). Should our mean field variational inference method underperform in predictive ability, we will pivot to using other values of α in the $(0, 1]$ interval (Yang, Pati, and Bhattacharya 2020). We will then assess performance in terms of predictive ability as a function of α.

Specific Aim 2: We will develop a Bayesian statistical model for polygenic risk scores based on SNP effect estimates and estimates for SNP-SNP interaction effects Introduction One possible reason for the modest predictive performance of current polygenic risk scores is their collective failure to account for SNP-SNP interactions in their statistical modeling. This oversight is partially due to the computing resources that are needed to accommodate not only SNP main effects, but the very large number of possible genome-wide SNP-SNP interactions. The objective of Specific Aim 2 is to develop polygenic risk score statistical methods that model both SNP main effects and SNP-SNP interactions. While the sheer number of SNP-SNP interactions across the genome may make it too computationally costly to use sampling-based inference methods like LDPred2 (Privé, Arbel, and Vilhjálmsson 2020), we anticipate that the gains in efficiency from use of variational inference will make it computationally feasible for us to incorporate modeling of SNP-SNP interactions into our polygenic risk scores. Upon completion of Specific Aim 2, it is our expectation that we will have created a computationally scalable and efficient method for constructing polygenic risk scores that models both SNP main effects and SNP-SNP interactions. We expect that our modeling of SNP-SNP interactions will lead to improved predictive performance of our method relative to current standard methods, such as LDpred2 (Privé, Arbel, and Vilhjálmsson 2020) and PRS-CS (Ge et al. 2019).

Research Design

Study Data

As in Specific Aim 1, we will use data from 362,320 UK Biobank subjects at 1,117,493 genetic markers. We will examine dozens of diseases for every subject and will ensure that we consider traits across the spectrum of SNP heritability values and traits with distinct patterns of genetic architectures.

Statistical modeling

Statistical inference methods

Assessing predictive performance

Like in our strategy for Specific Aim 1, we will measure predictive performance through area under the receiver operating characteristic curve. For Specific Aim 2, we need to quantify the anticipated gains in predictive performance from modeling SNP-SNP interactions. To do this, we will compare areas under the curve for our polygenic risk scores that omit SNP-SNP interactions (i.e., those from Specific Aim 1) to those that model SNP- SNP interactions for every disease of interest. We expect to see improved performances for the polygenic risk scores that model SNP-SNP interactions, and we expect that the size of the performance improvement to be greater for those traits with greater SNP heritability values.

Expected Outcomes, Potential Problems & Alternative Strategies

Our expected outcomes from Specific Aim 2 include a new polygenic risk score statistical method that accounts for not only SNP main effects, but also models SNP-SNP interactions. With this more comprehensive modeling of genetic effects, we expect that our method with SNP-SNP interactions will outperform, in terms of predictive ability, current state-of-the-art polygenic risk scores, since none of them model SNP-SNP interactions. Potential problems include the possibility that our modeling of SNP-SNP interactions doesn't improve predictive performance over that of polygenic risk scores that model only SNP main effects. Should our initial studies on a limited set of diseases from the UK Biobank not provide evidence that our modeling of SNP-SNP interactions improves predictive performance, we will expand our study to examine more diseases, especially those with high SNP heritability values, in the UK Biobank.

References

Bateson, W. 1909. *Mendel's Principles of Heredity*. Cambridge University Press.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe. 2017. "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association* 112 (518): 859–77.

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.

Consortium, International HapMap 3 et al. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52.

Demontis, Ditte, Raymond K Walters, Joanna Martin, Manuel Mattheisen, Thomas D Als, Esben Agerbo, Gísli Baldursson, et al. 2019. "Discovery of the First Genome-Wide Significant Risk Loci for Attention Deficit/Hyperactivity Disorder." *Nature Genetics* 51 (1): 63–75.

Fisher, Ronald A. 1919. "XV.—the Correlation Between Relatives on the Supposition of Mendelian Inheritance." *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52 (2): 399–433.

Ge, Tian, Chia-Yen Chen, Yang Ni, Yen-Chen Anne Feng, and Jordan W Smoller. 2019. "Polygenic Prediction via Bayesian Regression and Continuous Shrinkage Priors." *Nature Communications* 10 (1): 1776.

Lee, Kwan-Yeung, Kwong-Sak Leung, Suk Ling Ma, Hon Cheong So, Dan Huang, Nelson Leung-Sang Tang, and Man-Hon Wong. 2020. "Genome-Wide Search for SNP Interactions in GWAS Data: Algorithm, Feasibility, Replication Using Schizophrenia Datasets." *Frontiers in Genetics* 11: 1003.

Li, Pei, Maozu Guo, Chunyu Wang, Xiaoyan Liu, and Quan Zou. 2015. "An Overview of SNP Interactions in Genome-Wide Association Studies." *Briefings in Functional Genomics* 14 (2): 143–55.

Luciano, Michelle, Saskia P Hagenaars, Gail Davies, W David Hill, Toni-Kim Clarke, Masoud Shirali, Sarah E Harris, et al. 2018. “Association Analysis in over 329,000 Individuals Identifies 116 Independent Variants Influencing Neuroticism.” *Nature Genetics* 50 (1): 6–11.

Mullins, Niamh, Andreas J Forstner, Kevin S O’Connell, Brandon Coombes, Jonathan RI Coleman, Zhen Qiao, Thomas D Als, et al. 2021. “Genome-Wide Association Study of More Than 40,000 Bipolar Disorder Cases Provides New Insights into the Underlying Biology.” *Nature Genetics* 53 (6): 817–29.

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

Privé, Florian, Hugues Aschard, Andrey Ziyatdinov, and Michael GB Blum. 2018. “Efficient Analysis of Large-Scale Genome-Wide Data with Two r Packages: Bigstatsr and Bigsnpr.” *Bioinformatics* 34 (16): 2781–87.

Ray, Kolyan, Botond Szabo, and Gabriel Clara. 2020. “Spike and Slab Variational Bayes for High Dimensional Logistic Regression.” *Advances in Neural Information Processing Systems* 33: 14423–34.

Vilhjálmsson, Bjarni J, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, et al. 2015. “Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores.” *The American Journal of Human Genetics* 97 (4): 576–92.

Wray, Naomi R, Stephan Ripke, Manuel Mattheisen, Maciej Trzaskowski, Enda M Byrne, Abdel Abdellaoui, Mark J Adams, et al. 2018. “Genome-Wide Association Analyses Identify 44 Risk Variants and Refine the Genetic Architecture of Major Depression.” *Nature Genetics* 50 (5): 668–81.

Yang, Yun, Debdeep Pati, and Anirban Bhattacharya. 2020. “α-Variational Inference with Statistical Guarantees.” *Annals of Statistics*.

Zhang, Chun-Xia, Shuang Xu, and Jiang-She Zhang. 2019. “A Novel Variational Bayesian Method for Variable Selection in Logistic Regression Models.” *Computational Statistics & Data Analysis* 133: 1–19.