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.

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.

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.

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.

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.

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.

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.