SALUD: Scalable Applications of cLinical risk Utility and preDiction

Pankhuri Singhal
Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA 19104, USA
Email: singhalp@pennmedicine.upenn.edu

Yogasudha Veturi
The Pennsylvania State University
University Park, PA 16801, USA
Email: yzv101@psu.edu

Renae Judy
Department of Surgery, Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA 19104, USA
Email: renae.judy@pennmedicine.upenn.edu

Yoson Park
Internal Medicine Research Unit, Pfizer Inc
Boston, MA 02139, USA
Email: yoson.park@gmail.com

Marijana Vujkovic
Department of Genetics, Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA 19104, USA
Email: vujkovic@pennmedicine.upenn.edu

Olivia Veatch
Department of Psychiatry & Behavioral Sciences, and Molecular & Integrative Physiology at the University of Kansas Medical Center
Kansas, MO 66103, USA
Email: oveatch@kumc.edu

Rachel Kember
Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania
Philadelphia, PA 19104, USA
Email: rkember@pennmedicine.upenn.edu

Shefali Setia Verma
Department of Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania 19104, USA
Philadelphia, PA
Email: shefali.setiaverma@pennmedicine.upenn.edu

This PSB 2023 session discusses challenges in clinical implication and application of risk prediction models, which includes but is not limited to: implementation of risk models, responsible use of polygenic risk scores (PGS), and other risk prediction strategies. We focus on the development and use of new, scalable methods for harmonizing and refining risk prediction models

by incorporating genetic and non-genetic risk factors, applying new phenotyping strategies, and integrating clinical factors and biomarkers. Lastly, we will discuss innovation in expanding the utility of these prediction models to underrepresented populations. This session focuses on the overarching theme of enabling early diagnosis, and treatment and preventive measures related to complex diseases and comorbidities.

Keywords: Risk Prediction, risk factors, clinical implementation polygenic risk scores, complex human diseases

1. Introduction:

Genetic variants each harboring small phenotypic effects are shown to collectively contribute to complex trait and disease risk. Genome-wide association studies (GWAS), a mainstay of genetics research, are widely used to identify such common genetic variants (single nucleotide polymorphisms or SNPs) that convey increased or decreased risk for complex traits in populations. Due to the polygenic nature of complex traits, reliably predicting disease susceptibility or risk often requires studies of large sample sizes. To address this, large biobanks such as the Million Veteran Program (MVP) and UK Biobank, and consortia such as the Global Lipids Genetics Consortium (Graham et al., 2021), Global Biobank Meta-Analysis Initiative (Zhou et al., 2021), and Genetic Investigation of Anthropometric Traits (Yengo et al., 2018), among several others, have been successful at identifying and validating genetic components of complex traits based on sample sizes ranging from hundreds of thousands to over a million. Nevertheless, identifying people at risk of disease prior to the presentation of symptoms remains one of the main challenges and goals of precision medicine. Countless hours and resources are spent in understanding the pathophysiology of complex diseases and identifying clinical, genetic, and exposure risk factors that influence the risk of prevalent diseases that substantially impact public health such as breast cancer, coronary artery disease (CAD), obesity, and type 2 diabetes.

Consequently, estimating the disease risk of patients based on their common genetic variants by aggregating the weighted sum of the trait-affected alleles from GWAS into polygenic scores [PGS, also known as genetic risk scores (GRS) or polygenic risk scores (PRS)] has gained popularity (Wand et al., 2021). PGS provides an opportunity to estimate an individual's genetic risk (or predisposition) for complex diseases or traits. This is set as a non-modifiable lifetime risk and could be utilized prior to symptom onset to improve patients' health by predicting relatively modifiable factors such as lifestyle, nutrition, clinical, and other cumulative non-genetic risks that may act over multiple years (Torkamani et al., 2018). PGS capture a larger proportion of genetic liability than individual SNPs alone and have already been used to identify patients with disease risk equivalent to monogenic mutations, predict mortality, identify cases with earlier disease onset, and provide evidence for cross-trait associations. Recently, focus and interest have shifted from the theoretical application of PGS post hoc in large populations to the implementation of these methods for individual patients in clinical practices. Risk models such as BOADECIA for breast cancer (Lee et al., 2019) and cardioriskSCORE for CAD include PGS along with other clinical risk factors such as family history. Models for cancer risk have been integrated into wider gene screening panels such as PGLNext and ColoNext that test a subset of genes to provide cancer-type specific testing as a consumer product.

We are in a golden digital age for medicine in which individuals have access to their health records and genetic data at their fingertips. There is a strong public interest in better understanding personal genetics made clear in the various companies that have been founded in the last decade to bridge the gap between consumer and clinician. Companies like 23&Me provide genetic insight into trait and disease risks, while others focus on aspects of genetics including ancestry, embryo screening, fertility, cancer risk, allergy predispositions, diet optimization and weight loss, immune health, and cardiovascular event prediction. PGS have become a particular focus area of the health technology sector as a means of data-driven disease prevention. Numerous companies are geared towards providing genetics-based health risk predictions based on the application of PGS. These have been designed not only for the average individual but also for companies looking to build wellness incentive programs within their own businesses. Some PGS-focused companies provide risk score prediction as a clinical tool or platform for health systems and healthcare providers to implement in their clinics and hospitals. The wide scope of commercial applications underscores the keen interest in exploring genetic risk prediction. The direct-to-consumer model, however, comes with a great responsibility to critically examine the methodology with respect to health equity and diversity.

Despite recent advancements, a number of aspects of PGS require evaluation. PGS generated from currently available GWAS typically explain only a small proportion, 2-10%, of trait variation (Stringer et al., 2011). Moreover, a disproportionate majority (>78%) of participants in genetic studies are of European descent, limiting applications of PGS for many traits to individuals from this ancestry only (Sirugo et al., 2019). Also, many questions remain regarding best practices for the harmonization of multiple risk factors into clinically relevant models, particularly when including genetic factors in non-European populations or in longitudinal cumulative risk predictions.

Consented EHR-linked biobanks provide a vast and continuously growing repository of longitudinal data on diverse clinical populations that can fuel clinical, genetic, and epidemiologic research. Risk prediction models are not limited to a single phenotype or to a cross-sectional analysis of patient health. With the availability of multidimensional genomic and EHR data, longitudinal and time-series analyses can be conducted to investigate patient disease trajectories (Jensen et al., 2014). Complex genetic diseases often do not present phenotypically in the same way, in the same timeframe, in all patients (Woodward et al., 2022). Understanding which types of individuals develop certain conditions– and when– is essential for prognostics and disease prevention. Moreover, linking phenotypic patterns with genetic underpinnings can improve the predictive power of risk models. Such integrated risk prediction could be built upon a variety of machine learning methodologies and clinical and genomic data types. This is especially useful for understanding both the etiological basis for disease comorbidity and the architecture of disease co-occurrence (Monchka et al., 2022). Various network and statistical approaches have been applied to determine shared genetic components of comorbid conditions and the interactions between disease-associated gene products (Barabási et al., 2011). Leveraging longitudinal data in these analyses can provide a predictive aspect for disease onset. In addition, other kinds of omics data (e.g., transcriptomics, proteomics, metabolomics) can explain variance attributable to genetics as well as some lifestyle/environmental factors (Kim et al., 2015). Furthermore, the fact that EHR data are collected in real-world clinical settings makes them particularly valuable for research aimed at reflecting population diversity.

2. Overview of the contributions

The SALUD session keynote talk by Dr. Cooke-Bailey entitled "Pause, Reflect, Redirect: Clinical Scalability of Genetic Risk Scores Remains Limited due to Lack of Diversity" will focus on the utility of risk scores across disease, model, and scope of genetic data, as well as and what remains lacking across the breadth of these approaches in clinical scalability and broad applicability. While future GRS and PRS may serve as surrogate measures for disease risk, the current landscape leaves much room for improvement in clinical implementation across different ancestral groups. Key to realizing the true power of clinical and genetic risk models is intentional focus on improving representation of data from populations that have historically been underrepresented in research. This session will be focused on the utility of risk scores across several common and complex disorders as described briefly below.

One of the goals of precision medicine is to be able to stratify patients based on their genetic risk for a disease using GRS to inform future screening and intervention strategies. However, the variants used to calculate these scores are often based on European (EUR) ancestry individuals, limiting their clinical utility. Study titled "*Diversity is key for cross-ancestry transferability of glaucoma genetic risk scores in Hispanic Veterans in the Million Veteran Program*" by *Waksmunski et al.* addresses the challenges of applying GRS in complex conditions like primary open-angle glaucoma (POAG). POAG disproportionately affects individuals of African and Hispanic (HIS) ancestries. This study evaluates the risk stratification performance of POAG GRS based on cross-ancestry variants in EUR and HIS individuals.

Abdominal aortic aneurysms (AAA) are common enlargements of the abdominal aorta which can grow larger until rupture, often leading to death. Recent large-scale genome-wide association studies have identified genetic loci associated with AAA risk. Study titled "*Predictive models for abdominal aortic aneurysms using polygenic scores and PheWAS- derived risk factors"* by *Hellwege et al.* combines known risk factors, PRS, and precedent clinical diagnoses from electronic health records (EHR) to develop predictive models for AAA. The resulting models improve identification of people at risk of a AAA diagnosis compared with existing guidelines.

Study titled "*Quantifying factors that affect polygenic risk score performance across diverse ancestries and age groups for body mass index"* by *Hui and Xiao et al.* addresses the challenge of limited transferability of PRS across groups that differ in ancestry or sample characteristics. To evaluate these factors in the PRS generation process, the authors quantified the effects of ancestry, genome-wide association study summary statistics sample size, and LD reference panel on PRS performance. This was done using a cross-ancestry and age-specific approach. PRS for body mass index (BMI) was generated for this analysis. Furthermore, comorbidities and clinical associations in electronic health records with PRS for BMI were explored.

Late-onset Alzheimer's disease (LOAD) is a polygenic disorder with a long prodromal phase, making early diagnosis challenging. PRS leverage combined effects of many loci to predict LOAD risk, but often lack sensitivity to preclinical disease changes, limiting clinical utility. Study titled

"*Resilience polygenic risk score may be sensitive to preclinical disease changes"* by *Eissman et al*. generates a resilience phenotype to model better-than-expected cognition given LOAD biomarker levels in order to bolster preclinical polygenic risk prediction. The resulting LOAD PRS and resilience PRS models together are evaluated for prediction of preclinical disease status among dementia-free and biomarker-positive individuals.

3. Conclusion

Developing accurate risk prediction models for disease is one of the main goals of precision medicine. The addition of genetic data to these models could enhance their performance. However, there are many questions about appropriate implementation, interpretation, and derivation of genetic risk prediction models. The studies presented in this session explore these issues by combining genetic scores with known risk factors to test the improvement in performance, enhance transferability of genetic scores in diverse ancestries, and evaluate the ability of models including genetic scores to predict preclinical disease status. This research is essential as we move towards incorporating genetic risk prediction models in clinical practice.

4. Acknowledgements

PS is supported by F31AG069441-01.

References

1. Barabási, A. L., Gulbahce, N., & Loscalzo, J. (2011). Network medicine: A network-based approach to human disease. In *Nature Reviews Genetics* (Vol. 12, Issue 1, pp. 56–68). https://doi.org/10.1038/nrg2918
2. Graham, Sarah E., et al. "The power of genetic diversity in genome-wide association studies of lipids." *Nature* 600.7890 (2021): 675-679.
3. Jensen, A. B., Moseley, P. L., Oprea, T. I., Ellesøe, S. G., Eriksson, R., Schmock, H., Jensen, P. B., Jensen, L. J., & Brunak, S. (2014). Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms5022
4. Kim, D., Joung, J. G., Sohn, K. A., Shin, H., Park, Y. R., Ritchie, M. D., & Kim, J. H. (2015). Knowledge boosting: A graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction. *Journal of the American Medical Informatics Association*, *22*(1), 109–120. https://doi.org/10.1136/amiajnl-2013-002481
5. Lee, Andrew, et al. "BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors." *Genetics in Medicine* 21.8 (2019): 1708-1718.
6. Monchka, B. A., Leung, C. K., Nickel, N. C., & Lix, L. M. (2022). The effect of disease co-occurrence measurement on multimorbidity networks: a population-based study. *BMC Medical Research Methodology*, *22*(1). https://doi.org/10.1186/s12874-022-01607-8

7. Sirugo, G., Williams, S. M., & Tishkoff, S. A. (2019). The Missing Diversity in Human Genetic Studies. *Cell*, *177*(1), 26–31. https://doi.org/10.1016/j.cell.2019.02.048
8. Stringer, S., Wray, N. R., Kahn, R. S., & Derks, E. M. (2011). Underestimated effect sizes in GWAS: fundamental limitations of single SNP analysis for dichotomous phenotypes. *PloS One*, *6*(11), e27964. https://doi.org/10.1371/journal.pone.0027964
9. Torkamani, Ali, Nathan E. Wineinger, and Eric J. Topol. "The personal and clinical utility of polygenic risk scores." *Nature Reviews Genetics* 19.9 (2018): 581-590.
10. W and, H., Lambert, S. A., Tamburro, C., Iacocca, M. A., O'Sullivan, J. W., Sillari, C., Kullo, I. J., Rowley, R., Dron, J. S., Brockman, D., Venner, E., McCarthy, M. I., Antoniou, A. C., Easton, D. F., Hegele, R. A., Khera, A. v., Chatterjee, N., Kooperberg, C., Edwards, K., … Wojcik, G. L. (2021). Improving reporting standards for polygenic scores in risk prediction studies. In *Nature* (Vol. 591, Issue 7849, pp. 211–219). Nature Research. https://doi.org/10.1038/s41586-021-03243-6
11. Woodward, A. A., Urbanowicz, R. J., Naj, A. C., & Moore, J. H. (2022). Genetic heterogeneity: Challenges, impacts, and methods through an associative lens. In *Genetic Epidemiology*. John Wiley and Sons Inc. https://doi.org/10.1002/gepi.22497
12. Yengo, Loic, et al. "Meta-analysis of genome-wide association studies for height and body mass index in∼ 700000 individuals of European ancestry." *Human molecular genetics* 27.20 (2018): 3641-3649.
13. Zhou, Wei, and Global Biobank Meta-analysis Initiative. "Global Biobank Meta-analysis Initiative: Powering genetic discovery across human diseases." *medRxiv* (2021).