Overcoming health disparities in precision medicine

Francisco M. De La Vega,[1] Kathleen C. Barnes,[2] Keolu Fox,[3] Alexander Ioannidis,[4] Eimear Kenny,[5] Rasika A. Mathias,[6] and Bogdan Pasaniuc.[7]

[1]Tempus Labs, Inc.,[2]Galatea Bio, Inc., [3]University of California San Diego, [4]Stanford University School of Medicine, [5]Ichan School of Medicine at Mount Sinai, [6]Johns Hopkins School of Medicine, [7]University of California Los Angeles.

1. Overview

Precision medicine and precision public health rely on the premise that determinants of disease incidence and differences in response to interventions can be identified, and their biology can be understood well enough for the development of individualized interventions that reduce the risk of disease and improve treatment. At the same time, well-documented racial and ethnic disparities exist throughout healthcare at the patient, provider, and healthcare system levels. These disparities are driven by a complex interplay among social, psychosocial, lifestyle, environmental, health system, and biological determinants of health (Freedman, et al. 2021). The aim of the PSB 2024 session "Overcoming health disparities in precision medicine" is to elicit the development of new methods and concepts than can be used in uncovering undetected biases, develop effective therapies and fair AI to improve precision healthcare and help reduce these disparities, and ultimately improve health equity.

2. Dealing with the lack of diversity in current research datasets

An overwhelming focus on individuals of European descent in past genomic studies, which account for 86% of all such research, has created inequities in precision medical insights and has limited scientific discovery (Fatumo et al. 2022). It is imperative to diversify genomic research data and to make investments aimed at understanding and eliminating these health inequalities.

In the meantime, methods that can use the currently available genomic and clinical data, which admittedly are lacking in diversity, to provide equitable prediction of phenotypes are needed. The paper by Comajoan Cara et al. (2024) in this proceedings introduces PopGenAdapt, a model that tackles the lack of diversity in genomic datasets by using semi-supervised domain adaptation techniques. The model effectively leverages labeled data from individuals of European ancestry and both labeled and unlabeled data from underrepresented populations. When tested in populations from Nigeria, Sri Lanka, and Hawaii, PopGenAdapt showed significant improvement in predicting disease outcomes compared to existing methods, highlighting its potential for more inclusive biomedical research.

On the other hand, the paper by Bonet et al. (2024) introduces a machine learning toolkit designed to directly improve the accuracy of genomic-based medical predictions for

underrepresented populations. By employing techniques such as gradient boosting, ensembling, and population-conditional re-sampling techniques to address the lack of diversity, the method enhances phenotype prediction accuracy, achieving results comparable to those for well-represented European populations.

Ancestry can impact gene expression prediction methods as potentially undiscovered population variants and eQTNs affecting gene expression may exists in understudied populations. The paper by Mishra et al. introduces LA-GEM, a gene imputation model that incorporates local ancestry (LA) to improve gene expression predictions in African American populations. Tested on a cohort of 60 African American hepatocyte primary cultures, LA-GEM outperformed existing models like PrediXcan by reliably predicting the expression of unique genes critical to drug metabolism in this sample. The study highlights the value of leveraging local ancestry in gene imputation models for admixed populations to better understand disease susceptibility and drug response in all populations.

3. Development of fair machine learning algorithms

The development of fair algorithms and machine learning in healthcare is crucial for reducing health disparities, improving diagnostic accuracy, and building public trust. By minimizing biases, equitable healthcare and regulatory compliance is promoted, leading to more economically efficient systems.

One of the first steps to algorithmic equity is the thorough exploration of the input data to be used in their training for inherent and occult biases. The paper by Orlenko et al. (2024) provides examples of such necessary data exploration using cluster analysis to identify two distinct subgroups of elective spinal fusion patients based on insurance type. These findings reveal significant differences in characteristics and post-surgery outcomes related to socioeconomic and racial disparities. The aim is to inform the design of machine learning models to ensure fairness and minimize bias in healthcare predictions.

Methods designed to provide fair algorithmic predictions from the ground up are needed as well. Jun et al. (2024) use a fairness algorithm, Fairness-Aware Causal paThs (FACTS), to analyze nine years of electronic health records and social determinants of health to quantify disparities in MRSA infection outcomes. The study identified moderate disparities in age, gender, race, and income, revealing that comorbidities played a role in these disparities. Factors like kidney impairment and drug use affected racial disparity, while income and healthcare access affected gender disparity. The findings highlight the need for policies that address both clinical factors and social determinants to mitigate health disparities.

4. Race, genetic ancestry, and population structure

The persistent use of "race" and "ethnicity" in precision medicine, classifications rooted in perceived physical characteristics and cultural backgrounds, has generated substantial discussion (Nat. Acad. Sci. Eng. Med., 2023). However, for the purpose of examining health equity, these categories remain essential for identifying and addressing systemic disparities (Kahu et. Al. (2021). Established by the U.S. Office of Management and Budget in 1995, these classifications are integral for organizing

data on social determinants of health and population demographics. They guide resource allocation, policymaking, and the development of culturally sensitive healthcare interventions.

The paper by Rhead et al. (2024) tackles the problem of missing disaggregated race and ethnicity data in real-world databases by introducing methods for imputing these categories using genetic ancestry from available genetic data. Analyzing data from over 100,000 cancer patients, ancestry-based machine learning methods were shown to outperform existing race imputation algorithms based on geolocation and surnames commonly used in administrative health data. The research offers a new way to improve real-world healthcare data for studying and ensuring healthcare equity and to enable its use in the development of diversity plans for clinical trials soon to be required per FDA guidance.

On the other hand, the study by Seagle et al. (2024) analyzes the genetic ancestry of 35,842 individuals over 100 birth years in the Southeastern United States, finding increasing levels of genetic admixture and heterozygosity in younger populations since 1990. This rise in diversity poses challenges to traditional genotype-phenotype relationship studies. The researchers explore the impact of increased admixture on health outcomes, discovering that greater genetic diversity was associated with protective effects against female reproductive disorders but elevated risks for diseases linked to autoimmune dysfunction. This highlights the influence of ancestral complexity on health disparities.

The social construct of race and ethnicity is far from precise, serving as a poor proxy for ancestry. In this vein, the study by Piekos et al. (2024) employs genetic ancestry rather than race to assess disease risk factors, leveraging data from the BioVU biobank. Researchers estimated six ancestry proportions and performed phenome-wide association studies, finding varying risks for conditions like 'Neoplasms' and 'Pregnancy Complications' based on different ancestries. The study also found that linear modeling was sufficient for assessing hypertension and atrial fibrillation risk in relation to ancestry, but not for renal failure, indicating the need for more complex models in certain cases.

5. Conclusion

The increased attention to social justice has emphasized the urgent need to tackle health disparities more effectively. Advanced computational and statistical approaches are essential for assessing and mitigating these disparities in healthcare. Their adoption is not just a technological advancement but also an ethical necessity for creating a healthcare environment that serves all communities effectively. We believe that the new methods in the collection of research papers accepted to this PSB 2024 proceedings can contribute to overcome disparities in precision medicine.

6. Acknowledgments

We thank the anonymous reviewers that helped in the peer review process of the submissions to this session.

References

Bonet D., Levin M., Mas Montserrat D. and Ioannidis A.G. (2024) Machine Learning Strategies for Improved Phenotype Prediction in Underrepresented Populations. In *Pacific Symposium on Biocomputing 2024*.

Comajoan Cara M., Mas Montserrat D., Ioannidis A.G. (2024) PopGenAdapt: Semi-Supervised Domain Adaptation for Genotype-to-Phenotype Prediction in Underrepresented Populations. In *Pac. Symp. on Biocomputing*.

Jun I., Ser S., Cohen S., Xu J., Lucero R.J., Bian J. and Prosperi M. (2024) Quantifying Health Outcome Disparity in Invasive Methicillin-Resistant Staphylococcus aureus Infection using Fairness Algorithms on Real-World Data. In *Pac. Symp. on Biocomputing*.

Fatumo S., Chikowore T., Choudhury A., Ayub M., Martin A.R., Kuchenbaecker K. (2022) A roadmap to increase diversity in genomic studies. *Nat Med.* 2022 Feb;28(2):243-250.

Freedman J.A., Abo M.A., Allen T.A., Piwarski S.A., Wegermann K., and Patierno S.R. (2021) Biological Aspects of Cancer Health Disparities. *Ann. Rev. Med.* 72:229-241

Kahu T.J., Ghazal Read, J. and Scheitler, A.J. (2021) The Critical Role of Racial/Ethnic Data Disaggregation for Health Equity. *Pop. Res. Pol. Rev.* 40:1-7.

Mishra M., Nahlawi L., Zhong Y., De T., Yang G., Alarcon C. and Perera M.A. (2024) LA-GEM: imputation of gene expression with incorporation of Local Ancestry. In *Pac. Symp. on Biocomputing*.

National Academies of Sciences, Engineering, and Medicine, Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field*. (National Academies Press (US), (2023).

Orlenko A., Freda P.J., Ghosh A., Choi H., Matsumoto N., Bright T.J., Walker C.T., Obafemi-Ajayi T., and Moore J.H.. (2024) Cluster Analysis reveals Socioeconomic Disparities among Elective Spine Surgery Patients. In *Pac. Symp. on Biocomputing*.

Piekos J.A., Kim J., Keaton J.M, Hellwege J.N., Velez Edwards D.R., and Edwards T.L. (2024) Evaluating the Relationships Between Genetic Ancestry and the Clinical Phenome. In *Pac. Symp. on Biocomputing*.

Rhead B., Haffener P.E., Pouliot Y. and De La Vega F.M. (2024) Imputation of race and ethnicity categories using continental genetic ancestry from real- world genomic testing data. In *Pac. Symp. on Biocomputing*.

Seagle H.M., Mautz B.S., Hellwege J.N., Li C., Xu Y., Zhang S., Roden D.M., McGregor T.L., Velez Edwards D.R., Edwards T.L. (2024) Evidence of recent and ongoing admixture in the U.S. and influences on health and disparities. In *Pac. Symp. on Biocomputing*.