Subtyping Social Determinants of Health in *All of Us*: Opportunities and Challenges for Designing Precision Interventions

Suresh K. Bhavnani, Ph.D., M.Arch.[1,2] Weibin Zhang, Ph.D.,[1] Daniel Bao, B.S.,[1] Yong-Fang Kuo, Ph.D.[1], Susanne Schmidt, Ph.D.,[3] Monique R. Pappadis, Ph.D., M.Ed., FACRM,[1] Alex Bokov, Ph.D.,[4]Timothy Reistetter, Ph.D., OTR.,[5] Shyam Visweswaran*, M.D., Ph.D.,[7,8] Brian Downer*, Ph.D.[1] (* shared senior authorship)

[1]School of Public and Population Health, [2]Inst. for Translational Sciences, Univ. of Texas Medical Branch, Galveston, TX, USA; [3]Dept. of Population Health Sciences, Long School of Medicine, [4]School of Health Professions, Univ. of Texas Health San Antonio, TX, USA; [5]Dept. of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA.

Introduction

Social determinants of health (SDoH), such as financial resources[1] and housing stability,[2] can affect between 30-55% of people's health outcomes.[3] While many studies[4] have identified strong associations among specific SDoH and adverse health outcomes, most people have multiple SDoH that impact their daily lives,[5] which can co-occur to form SDoH subtypes, critical for designing targeted interventions. However, analysis of such subtypes requires the integration of personal, clinical, social, and environmental variables from a large cohort, which is only now becoming possible through the *All of Us* Research Program (*All of Us*).[6]

The data in *All of Us* provides an unprecedented opportunity to transform our understanding of SDoH. This program aims to collect health surveys including a wide range of SDoH variables, electronic health records (EHRs), whole sequence genome data, physical measurements, and personal digital information from one million or more individuals. Furthermore, the program focuses on including data from populations that have been traditionally underrepresented in biomedical research.[6] However, little is known about the range and response of SDoH in *All of Us*, and how they co-occur to form subtypes, which are critical for designing targeted interventions. To address these gaps, we characterized a wide range of SDoH (d=110) across the full *All of Us* cohort (n=372,397), and used them to identify and interpret SDoH subtypes through the use of scalable and generalizable machine learning methods (see preprint[7] for full study).

Method

Research Questions. Question-1: What is the range and response to survey questions related to SDoH? Question-2: How do SDoH co-occur to form subtypes, and what are their risk for adverse health outcomes?

Methods. For Question-1, 3 SDoH experts reviewed all 1,113 questions across 7 *All of Us* non-COVID health surveys. Through consensus, they identified 110 questions relevant to SDoH from 4 surveys *(The Basics, Overall Health, Health Care Access & Utilization, and SDoH*). These 110 survey questions spanned the full range of the 5 SDoH domains identified by *Healthy People 2030* (HP-30).[8] However, due to the uneven granularity among these questions (e.g., *cannot afford dental care*, and *cannot afford prescriptions* had finer granularity compared to *single household*), the experts recommended that the 110 SDoH questions be grouped into 18 SDoH subdomains with consistent granularity and therefore higher clinical interpretability. A participant was defined as having an SDoH subdomain if they had a valid response (no "skip" or "choose not to answer") to $\geq$1 of the SDoH questions grouped within that subdomain. The responses to the 110 SDoH questions, and the SDoH subdomains were characterized across the full *All of Us* cohort (n=372,397, V6).

For Question-2, due to the systematic missingness in survey responses, we identified all participants with valid responses to the 110 SDoH questions, and randomly divided them into training and replication datasets. We used bipartite network analysis[9] to identify SDoH subtypes using the following steps: (1) represented the data as a bipartite network where nodes consisted of participants or SDoH subdomains, and edges connecting them were weighted using inverse probability weighting (IPW)[10] to rebalance the demographic proportions in our sample, compared with the full cohort; (2) used *bicluster modularity maximization*[9] to automatically identify the number and members of participant-SDoH subdomain biclusters, and measured modularity (Q) or the quality of biclustering; (3) measured the significance of Q by comparing it to a distribution of Q generated from 1000 random permutations of the network; (4) visualized the results using force-directed algorithms; (5) repeated the analysis using the replication dataset; and (6) used the Rand Index (RI)[11] to measure the degree of similarity in SDoH subdomain co-occurrence in the training and replication datasets, and measured the significance of RI by comparing it to a distribution of RI generated from random permutations of the training and the replication datasets. Furthermore, we used multivariable logistic regression to measure the odds ratio (OR) of participants in each bicluster compared with the other biclusters to estimate their risk for three outcomes known to be impacted by SDoH barriers (depression, delayed medical care, emergency room visits in the last year), using demographics as covariates, and corrected for multiple testing. Finally, we asked 3 domain experts to independently infer the subtype labels, in addition to the potential mechanisms that precipitate their adverse health outcomes and interventions to prevent them, and arrive at a consensus for their interpretations.

Results. For Question-1, we identified 110 SDoH questions across 4 surveys, which were categorized into 18 SDoH subdomains, and covered all 5 domains in HP-30. However, the results also revealed a large degree of missingness in survey

responses (1.76%-84.56%), with later surveys having significantly fewer responses compared to earlier ones, and significant differences in race, ethnicity, and age of participants when compared to the full cohort. For Question-2, the subtype analysis (training n=12,913, d=18) identified 4 biclusters with significant biclusteredness (Q=0.13, random-Q=0.11, z=7.5, P<0.001), and significant replication (Real-RI=0.88, Random-RI=0.62, P<.001). Furthermore, there were significant associations between specific subtypes and the outcomes, each with meaningful interpretations and potential precision interventions. For example, the subtype *Socioeconomic Barriers* included the SDoH subdomains *employment, food security, housing, income, literacy*, and *education attainment*, and had a significantly higher odds ratio (OR=4.2, CI=3.5-5.1, P-corr<.001) for depression, when compared to the subtype *Sociocultural Barriers*. As inferred by the SDoH experts, individuals that match this subtype profile could be screened early for depression and referred to social services for addressing combinations of SDoH such as housing and income. Finally, the identified subtypes spanned one or more of the 5 HP-30 SDoH domains[8] revealing the difference between the current knowledge-based SDoH domains, and the above data-driven subtypes. For example, the subtype *Socioeconomic Barriers* spanned both *Economic Stability* and *Education Access and Quality*, reflecting the complexity of how SDoH co-occur in the real world, and their potential use in the design of models to predict adverse health outcomes, and the design of interventions.

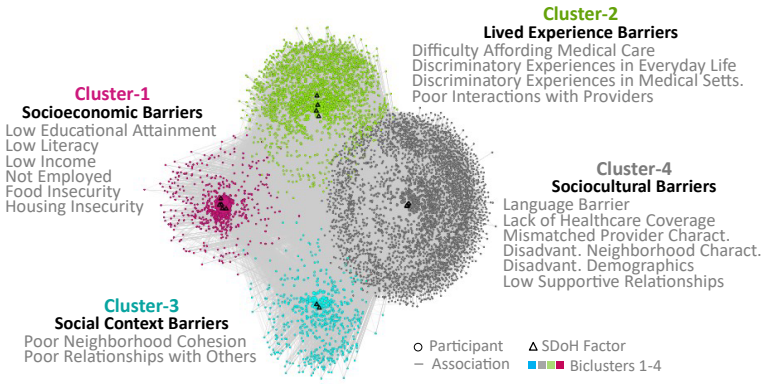

Fig. 1. Four subtypes in the training dataset consisting of subgroups of participants (n=6492), and their most frequently co-occurring SDoH subdomains (d=18).

Conclusions and Future Research

The results revealed translational and theoretical implications. From a **translational perspective**, the results revealed that the SDoH subtypes not only had statistically significant clustering and replicability, but also had significant associations with adverse health outcomes such as depression, with direct implications for designing targeted SDoH interventions. From a **theoretical perspective**, these SDoH subtypes spanned multiple SDoH domains defined by HP-30[8] revealing the complexity of SDoH in the real-world, and aligning with influential SDoH conceptual models such as by Dahlgren-Whitehead.[5] Furthermore, our subtyping code currently on the *All of Us* workbench consists of generalizable and scalable machine learning methods that can be used to periodically rerun the analysis as the *All of Us* cohort continues to evolve. Our future research will integrate other datatypes into the analysis including genomic information to enable the design of precision interventions, and to develop models for predicting adverse outcomes which incorporate membership into one or more of the SDoH subtypes.

Acknowledgements. Funded in part by UTMB Cancer Center, and UTMB CTSA (UL1 TR001439).

References

1. Weida EB, Phojanakong P, Patel F, Chilton M. Financial health as a measurable social determinant of health. *PloS one.* 2020;15(5):e0233359.
2. Kushel MB, Gupta R, Gee L, Haas JS. Housing instability and food insecurity as barriers to health care among low-income americans. *Journal of general internal medicine.* 2006;21(1):71-77.
3. WHO. Social determinants of health. https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1.
4. Lucyk K, McLaren L. Taking stock of the social determinants of health: A scoping review. *PloS one.* 2017;12(5):e0177306.
5. Dahlgren G, Whitehead M. The Dahlgren-Whitehead model of health determinants: 30 years on and still chasing rainbows. *Public health.* 2021;199:20-24.
6. Denny JC, Rutter JL, Goldstein DB, et al. The "All of Us" Research Program. *The New England journal of medicine.* 2019;381(7):668-676.
7. Bhavnani S, Zhang W, Bao D, et al. Subtyping Social Determinants of Health in All of Us: Opportunities and Challenges in Integrating Multiple Datatypes for Precision Medicine. *MedRxiv (preprint)* 2023; https://www.medrxiv.org/content/10.1101/2023.01.27.23285125v2.full.pdf.
8. Health.gov. Social Determinants of Health (Healthly People 2030). 2022; https://health.gov/healthypeople/priority-areas/social-determinants-health. Accessed 2/28/2023, 2023.
9. Barber MJ. Modularity and community detection in bipartite networks. *Physical Review E.* 2007;76(6):066102.
10. Thoemmes F, Ong AD. A Primer on Inverse Probability of Treatment Weighting and Marginal Structural Models. *Emerging Adulthood.* 2016;4(1):40-59.
11. Rand WM. Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association.* 1971;66(336):846-850.