Subtyping Social Determinants of Health in All of Us:

Opportunities and Challenges in Integrating Multiple Datatypes for Precision Medicine

Suresh K. Bhavnani, Ph.D., M.Arch.,¹,²§ Weibin Zhang, Ph.D.,¹ Daniel Bao, B.S.,¹ Mukaila Raji, M.D., M.S., F.A.C.P.,³ Yong-Fang Kuo, Ph.D.,¹ Susanne Schmidt, Ph.D.,⁴ Monique R. Pappadis, Ph.D., MEd, FACRM,¹ Alex Bokov, Ph.D.,⁴ Timothy Reistetter, Ph.D., OTR.,⁵ Shyam Visweswaran*, M.D., Ph.D.,⁶,⁷ Brian Downer*, Ph.D.¹

¹School of Public and Population Health, University of Texas Medical Branch, Galveston, TX, USA
²Institute for Translational Sciences, University of Texas Medical Branch, Galveston, TX, USA
³Division of Geriatric Medicine, Department of Internal Medicine, University of Texas Medical Branch, Galveston, TX, USA
⁴Department of Population Health Sciences, Long School of Medicine, University of Texas Health San Antonio, San Antonio, TX, USA
⁵School of Health Professions, University of Texas Health San Antonio, San Antonio, TX, USA
⁶Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
⁷Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA

§Corresponding author
Suresh K. Bhavnani, Ph.D., M.Arch., FAMIA
School of Public and Population Health
Department of Biostatistics and Data Science
Institute for Translational Sciences
University of Texas Medical Branch
301 University Blvd
Galveston, TX, USA
email: subhavna@utmb.edu

* Shyam Visweswaran and Brian Downer share senior authorship

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
A. ABSTRACT

Project Background. Social determinants of health (SDoH), such as unstable employment during the pandemic, account for between 30-55% of people's health outcomes. While many studies have identified strong associations among specific SDoH and health outcomes, most people experience multiple SDoH in their daily lives. Analysis of this complexity requires the integration of personal, clinical, social, and environmental information from a large cohort of underrepresented populations, which is only recently being made available through the All of Us research program. 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 precision medicine interventions.

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

Methods. For Question-1, we characterized the range of SDoH questions across the surveys, and analyzed their responses. For Question-2, we used the following steps: (1) due to the missingness across the surveys, selected all participants with valid and complete SDoH data, and used inverse probability weighting to adjust their imbalance in demographics, compared to the full cohort; (2) asked domain experts to map the SDoH questions to SDoH subdomains, for enabling a more consistent granularity; (3) used bipartite modularity maximization to identify SDoH biclusters, their significance, and their replicability; (4) measured the association of each bicluster with 3 outcomes (depression, delayed medical care, emergency room visits in the last year) using multiple data types (surveys, electronic health records, and zip codes mapped to Medicaid expansion states), and (5) asked 3 domain experts to infer the subtype labels, their mechanisms, and potential targeted interventions.

Results. For Question-1, we identified 110 SDoH questions across 4 surveys, categorized into 18 SDoH subdomains covering all 5 domains in Healthy People 2030 (HP-30). However, there was 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 (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 of specific subtypes with the outcomes and with Medicaid expansion, 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. Individuals that match this subtype profile could be screened early for depression and referred to social services to address combinations of SDoH such as housing and income. Finally, the identified subtypes spanned one or more HP-30 domains, revealing the difference between the current knowledge-based SDoH domains, versus the data-driven subtypes, reflecting the complexity of how SDoH co-occur in the real world, and their potential use in designing interventions.

Community Impact. While several SDoH models including the Dahlgren-Whitehead conceptual model have identified SDoH domains, they have emphasized that real-world SDoH span multiple domains with complex interactions and feedback loops. However, this phenomenon has been difficult to analyze given the lack of large cohorts with underrepresented populations characterized by a wide range of SDoH and datatypes. The results from analyzing SDoH using the All of Us cohort provided direct evidence for this real-world phenomenon by showing that data-driven SDoH subtypes span one or more of the SDoH domains defined by Healthy People 2030. This result provides testable hypotheses in future studies that SDoH models based on data-driven subtypes will be more accurate and interpretable for predicting adverse health outcomes, when compared to existing models that use the knowledge-driven domains. Furthermore, the characterization of the range and response to SDoH across the entire All of Us cohort using over one hundred SDoH, should enable researchers to use the approach for characterizing other cohorts for identifying and addressing missingness. Finally, our workbench which focuses on subtyping SDoH, provides generalizable and scalable machine learning methods that can be used to periodically rerun the analysis as the All of Us cohort continues to evolve.
B. Introduction

Social determinants of health (SDoH), such as unstable employment during the pandemic, account for between 30-55% of people's health outcomes. While many studies have identified strong associations among specific SDoH and health outcomes, most people experience multiple SDoH concurrently in their daily lives. For example, limited access to education, unstable employment, and lack of access to healthcare tend to frequently co-occur across individuals leading to long-term stress and depression. Such complex interactions among multiple SDoH make it critical to analyze combinations of SDoH versus single factors. However, analysis of such co-occurrences and their risks for adverse health outcomes requires the integration of personal, clinical, social, and environmental information, critical for designing cost-effective and precision interventions. Unfortunately, the lack of databases containing such multiple datatypes from the same individuals has resulted in a fragmented understanding of how SDoH co-occur and impact health, critical for designing precision interventions.

The All of Us program provides an unprecedented opportunity to address this fragmented view of SDoH. This program aims to collect multiple data types from one million or more individuals with a focus on populations that have been traditionally underrepresented in biomedical research. These data types include electronic health records (EHRs), health surveys, whole sequence genome data, physical measurements, and personal digital information. Critically, All of Us provides several survey modules containing a wide range of SDoH, which in combination with other datatypes, could transform our understanding of high-risk combinations of SDoH.

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 precision medicine interventions. To address these gaps, we characterized more than a hundred SDoH in All of Us, which guided the methods we used to analyze how they co-occur to form subtypes, and their risk for health outcomes. The results helped to highlight the opportunities and challenges for conducting subtype analysis in All of Us, which integrates multiple datatypes through the use of scalable and generalizable machine learning methods targeted to precision medicine.

C. Background

Social Determinants of Health

The World Health Organization (WHO) defines SDoH as the “non-medical factors that influence health outcomes.” Specifically, these include the conditions in which people are born, grow, work, live, and age. Furthermore, such conditions are shaped by a wider set of forces such as economic and social policies, and systems such as discriminatory laws and structural racism.

Several models have proposed the factors and mechanisms involved in SDoH. These models were motivated by the concept of social gradient, an empirical phenomenon observed within and across nations, consistently showing that the lower an individual's socioeconomic position, the worse their health. To help explain the factors underlying the social gradient, the Dahlgren-Whitehead model proposed several interconnected layers of social determinants that influence health. As shown in Fig. 1, the innermost layer contains demographic and genetic factors which are largely unmodifiable. In contrast, the outer layers are modifiable to different degrees such as lifestyle (e.g., exercise and smoking), social and community networks (e.g., contact with supportive friends and family), living and working conditions (e.g., access to health care and employment), and broader socio-economic, cultural, and environmental conditions (e.g., crime in the neighborhood). While this model was not intended to provide explicit testable hypotheses, the factors within each layer are expected to co-occur and impact each other, in addition to responding to external forces such as racism, and capitalism when it is focused on financial profits at the expense of societal benefits.

These early SDoH models motivated numerous studies that analyzed associations among specific SDoH (e.g., immigration status and home density), their association with health outcomes (e.g., education and mortality), and how they manifest within subpopulations (e.g., patients with diabetes). More recently, organizations such as Centers for Disease Control and Prevention (CDC) and Healthy People 2030 have organized these empirical results into SDoH domains that roughly map to the Dahlgren-Whitehead model. For example, Healthy People...
2030 organizes SDoH empirical studies into five SDoH domains: (1) Economic Stability; (2) Education Access and Quality; (3) Health Care Access and Quality; (4) Neighborhood and Built Environment; and (5) Social and Community Context. Furthermore, the PhenX program (that provides well-established measurement protocols for use in biomedical and translational research) has identified SDoH protocols to enable more systematic data collection and analysis.18-20

While the above findings and categorizations have greatly improved our understanding of SDoH and their impact on health, they have been mostly analyzed based on snapshots of associations among a few factors and health outcomes. In contrast, SDoH models and recent empirical studies suggest that multiple SDoH tend to co-occur and impact each other. For example, during the pandemic, Hispanic and Black or African American individuals not only had a higher exposure to COVID-19 infections due to their front-line jobs and overcrowded living conditions, but also had a higher risk for serious infections due to prior conditions not addressed due to lack of healthcare access.2 Similarly, undocumented immigrants with lower incomes living in neighborhoods with high pollution, combined with the stress of deportation, have an increased risk of multiple chronic conditions such as depression and lung cancer.5 Such studies have resulted in the Centers for Medicare and Medicaid Services (CMS) emphasizing that SDoH are a multi-level construct which includes both individual and contextual factors that have complex interactions.21

The above co-occurrences of multiple SDoH and their impact on health directly reflect the interconnected layers of the Dahlgren-Whitehead shown in Fig. 1. However, analysis of such co-occurrences and their health outcomes requires large datasets with multiple datatypes that have only recently been made available through the All of Us program.

All of Us: Multiple Datatypes Across a Large Cohort of Underrepresented Americans

The All of Us research program7-9 (All of Us), sponsored by the Obama administration since 2015, aims to accelerate biomedical research to enable discoveries leading to individualized and equitable prevention and treatment. Such research is currently hampered due to the limited range of personal, clinical, social, and environmental variables available for the same individuals, limited representation in research datasets of socially marginalized populations, and limited access to individual-level data due to privacy laws.

To overcome these hurdles, All of Us provides three critical features: (1) a data repository that is projected to contain one million or more participants, with multiple datatypes including electronic health records (EHRs), health surveys, whole sequence genomic data, physical measurements, and personal digital information such as from Fitbits; (2) a cohort targeted to include 75% participants from populations underrepresented in research (race, ethnicity, gender, sex, sexual orientation, and disability) oversampled from the US population; and (3) strictly-enforced rules to prevent reidentification of participants by disallowing the download of any participant data, or reporting research results for subgroups less than 20. These rules allow analysis of the All of Us data to be categorized as non-human subjects research, which combined with training and personal authentication by researchers, has resulted in a substantial reduction in administrative hurdles.

As of 12/30/22 (Controlled Tier, version 6), All of Us contained 372,397 total participants, with 8.6% who had attempted all 9 health surveys (7 related to demographics and general health, and 2 related to COVID-19), and 26.5% who had genomic data. Critical to the current study is the recent addition of a survey specifically targeted to SDoH questions, which has been attempted by 15.5% in the All of Us cohort. A preliminary analysis revealed that SDoH appear to be distributed across multiple health surveys and EHR codes, with participants providing those data at different times on a rolling basis. However, little is known about the range and response of SDoH in All of Us, and how they co-occur to form subtypes, a critical step for selecting the methods to identify and interpret SDoH subtypes.

Computational Methods to Identify and Interpret Subtypes

A wide range of studies22-30 on topics ranging from molecular to environmental determinants of health have shown that most humans tend to share a subset of characteristics (e.g., comorbidities, symptoms, genetic variants), forming distinct subtypes (also referred to as subgroups or subphenotypes depending on the condition and variables analyzed). A primary goal of precision medicine is to identify such subtypes and infer their underlying disease processes to design interventions targeted to those processes.23,31 Methods to identify subtypes include: (a) investigator-selected variables such as race for developing hierarchical regression models,32 or assigning patients to different arms of a clinical trial, (b) existing classification systems such as the
Medicare Severity-Diagnosis Related Group (MS-DRG)\textsuperscript{33} to assign patients into a disease category for purposes of billing, and (c) computational methods such as classification\textsuperscript{34-36} and clustering\textsuperscript{26,37} to discover subtypes from data.

Several studies have used computational methods to identify subtypes, each with critical trade-offs. Some studies have used \textit{combinatorial} approaches\textsuperscript{38} (identify all pairs, all triples etc.), which are intuitive, but which can lead to a combinatorial explosion (e.g., enumerating combinations of the 31 Elixhauser comorbidities would lead to \textit{2}\textsuperscript{31} or 2147483648 combinations), with most combinations that do not incorporate the full range of symptoms (e.g., the most frequent pair of symptoms ignores what other symptoms exist in the profile of patients with that pair). Other studies have used \textit{unipartite} clustering methods\textsuperscript{36,37} (clustering patients or comorbidities, but not both together) such as k-means, and hierarchical clustering; and dimensionality-reduction methods such as principal component analysis (PCA) to help identify clusters of frequently co-occurring comorbidities.\textsuperscript{38-44}

However, such methods have well-known limitations including the requirement of inputting user-selected parameters (e.g., similarity measures, and the number of expected clusters), in addition to the lack of a quantitative measure to describe the quality of the clustering (critical for measuring the statistical significance of the clustering). Furthermore, because these methods are unipartite, there is no agreed-upon method to identify the patient subgroup defined by a cluster of variables, and vice-versa.

More recently, bipartite network analysis\textsuperscript{45} (see Appendix A for additional details) has been used to address the above limitations by automatically identifying \textit{biclusters}, consisting of patients and characteristics simultaneously. This method takes as input any dataset such as All of Us participants and their SDoH, and outputs a quantitative and visual description of biclusters (containing both participant subgroups, and their frequently co-occurring SDoH). The quantitative output generates the number, size, and statistical significance of the biclusters,\textsuperscript{46-48} and the visual output displays the quantitative information of the biclusters through a network visualization.\textsuperscript{49-51}

Bipartite network analysis therefore enables (1) the automatic identification of biclusters and their significance, and (2) the visualization of the biclusters critical for their clinical interpretability. Furthermore, the attributes of participants in a subgroup can be used to measure the subgroup risk for an adverse outcome, to develop classifiers for classifying a new participant into one or more of the subgroups, and to develop a predictive model that uses that subgroup membership for measuring the risk of an adverse outcome for the classified participant.

However, while several studies\textsuperscript{48,52-59} have demonstrated the usefulness of bipartite networks for the identification and clinical interpretation of subgroups, there has been no systematic attempt to identify SDoH subtypes mainly because of the lack of large cohorts containing a wide coverage of SDoH. The All of Us program provides an opportunity to use bipartite networks for the identification and interpretation of SDoH subtypes using a wide range of variables in a large cohort, and for analyzing their risk for health outcomes, a critical step in advancing precision medicine.

D. Method

Research Questions

Our analysis was guided by two research questions targeting the All of Us dataset:

1. \textit{What is the range and response to survey questions related to SDoH?}

2. \textit{How do SDoH co-occur to form subtypes, and what are their risk for adverse health outcomes?}

Data Description

\textit{Study Population.} In Question-1, we analyzed the full All of Us cohort (n=372,397) and characterized their responses to all the SDoH identified by the domain experts (described in the Variables subsection). For Question-2, we analyzed all participants (n=12,913) that had valid responses to the SDoH identified in Question-1, and used them to identify subtypes, and their risks for specific outcomes.

\textit{Variables.} For Question-1, 3 SDoH domain experts were asked to review all 1113 questions across 7 All of Us non-COVID health surveys, each of which is attempted once per participant (\textit{The Basics, Lifestyle, The Basics, Personal Medical History, Health Care Access \& Utilization, Family Health History, and SDoH}), and the 2843 Systematized Nomenclature of Medicine (SNOMED) codes related to SDoH.\textsuperscript{60} The domain experts arrived at a consensus for the SDoH across the surveys and the SNOMED codes. As the SDoH-related SNOMED codes in the EHR had very low usage (see Appendix B for a characterization), they were not further characterized.
In Question-2, to identify SDoH subtypes, we used (a) the SDoH identified from Question-1, (b) covariates including 3-digit zip code (to determine if participants in each subtype came from a state that accepted Medicaid expansion providing greater access to health insurance), stress, variants of the gene SLC6A4 (known to be highly prevalent in families with depression), and demographics (age, sex, race). Outcomes included the following: (1) Depression was selected as it is a common health outcome when individuals encounter SDoH in their daily lives such as long-term stress resulting from racism, and dysregulation of the hypothalamic-pituitary-adrenal axis (HPA) axis. Depression was defined as having a positive response to both of the following questions in the The Basics survey (“Are you still seeing a doctor or health care provider for depression?” and “Has a doctor or health care provider ever told you that you have Depression?”) or having SNOMED codes related to depression Codes in their EHR (35489007, 36923009, 370143000, 191616006, or 66344007), (2) Delayed Medical Care was selected as it often results from the lack of medical insurance, which can impact the use of medical care when needed leading to poorer health outcomes. Delayed medical care was defined as having one or more positive responses to 9 survey questions (delayed care due to: transportation, rural, nervousness, work, childcare, copay, elderly care, out of pocket, and deductible) from the Health Care Access & Utilization survey. (3) Emergency Room (ER) Visits in Last Year was selected because lack of medical insurance often results in individuals not seeking early medical care when needed, leading to an exacerbation of conditions precipitating one or more ER visits. As the survey questions that we used for SDoH subtyping were based on outcomes in the past year, we defined ER visits for a participant as having one or more ER visits (CPT 99281-99285) one year preceding the date when the SDoH survey was completed.

**Analytical Approach**

**Question-1:** What is the range and response to survey questions related to SDoH?

To address this question, we characterized all SDoH in *All of Us* at two levels of granularity: (1) SDoH questions based on the surveys used to collect the data, and (2) SDoH subdomains, which were groupings of the SDoH questions to form a coarser grained classification (see Table-1 which explains the difference between SDoH questions, domains, subdomains, and subtypes). These two levels of granularity of SDoH in *All of Us* were characterized as follows:

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>SDoH Domains</td>
<td>Five domains defined by health.gov</td>
<td>Economic Stability</td>
</tr>
<tr>
<td>SDoH Subdomains</td>
<td>19 subdomains present in All of Us</td>
<td>Food Security</td>
</tr>
<tr>
<td>SDoH Questions</td>
<td>SDoH related questions in All of Us surveys (categorized into SDoH domains and SDoH subdomains using domain knowledge)</td>
<td>How often do you have someone to help you read health-related materials?</td>
</tr>
<tr>
<td>SDoH Subtypes</td>
<td>Data-driven clusters of SDoH subdomains identified through bipartite network analysis</td>
<td>Cluster Members: Food Security, Educational Attainment, Employment, Housing, Income, and Literacy</td>
</tr>
</tbody>
</table>

**Table 1. Definitions for SDoH domains, subdomains, questions, and subtypes.**

**A. Identification and Coding of SDoH Questions in All of Us.** Three SDoH experts independently used their domain knowledge about SDoH to identify and code the SDoH questions, and to examine their range with respect to the five Healthy People 2030 domains using the following steps: (1) review all 1113 questions across 7 health surveys (excluding 2 related to COVID-19), and extract all SDoH questions that were relevant; (2) transform all positive or value-free questions into negative phrases and abbreviate them for interpretability in the graphs (e.g., “How often do you have someone help you read health-related materials?” was changed into “No one to help read health materials”); (3) reverse code, and dichotomize the abbreviated SDoH questions (e.g., Always/Often=1, and Never/Occasionally/Sometimes=0); and (4) categorize the SDoH questions into one of the five Healthy People 2030 SDoH domains (Economic Stability, Education Access and Quality, Health Care Access and Quality, Neighbourhood and Built Environment, and Social and Community Context). The three domain experts subsequently met and collaboratively resolved any differences between their coding schemes to arrive at a consensus.

**B. Identification and Coding of SDoH Subdomains of Questions in All of Us.** The same three SDoH experts arrived at a consensus to categorize one or more of the above SDoH questions in *All of Us*, into SDoH subdomains, and to examine their range using the following steps: (1) review the subgrouping labels of questions in the *All of Us* surveys, and integrate them to classify the SDoH into subdomains; (2) code a participant as having a “1” for a SDoH subdomain if they had one or more of the questions within that subdomain which had been answered with a “1”; and (3) categorize the SDoH subdomains into one of the five Healthy People 2030 domains using the following steps: (1) review all 1113 questions across 7 health surveys (excluding 2 related to COVID-19), and extract all SDoH questions that were relevant; (2) transform all positive or value-free questions into negative phrases and abbreviate them for interpretability in the graphs (e.g., “How often do you have someone help you read health-related materials?” was changed into “No one to help read health materials”); (3) reverse code, and dichotomize the abbreviated SDoH questions (e.g., Always/Often=1, and Never/Occasionally/Sometimes=0); and (4) categorize the SDoH questions into one of the five Healthy People 2030 SDoH domains (Economic Stability, Education Access and Quality, Health Care Access and Quality, Neighbourhood and Built Environment, and Social and Community Context). The three domain experts subsequently met and collaboratively resolved any differences between their coding schemes to arrive at a consensus.

Range and Responses to SDoH Questions and Subdomains

The above knowledge-based classification of SDoH questions and SDoH subdomains were analyzed to examine their range (with respect to the five Healthy People 2030 domains), and their response (across all participants in All of Us), using the following methods. (1) Bar graph displaying the number of participants that had valid answers (all responses other than “skip” or “choose not to answer”) to each of the SDoH questions, sorted by survey based on mean response, and then sorted by raw response within each survey. Additionally, to examine their range, each SDoH question/subdomain was colored by one of the five SDoH domains defined by Healthy People 2030. (2) Venn diagram showing how many participants had cross-sectionally valid responses to all identified SDoH questions/subdomains. (3) Table describing the number and proportion of race, ethnicity, sex, gender, and age between those that answered the SDoH questions/subdomains, versus those that did not have valid responses. (4) Frequency distribution of the number of SDoH questions/subdomains across participants that had valid responses for all the SDoH questions. The above plots are shown in the Results section.

Question-2: How do SDoH co-occur to form subtypes, and what are their associations with covariates and risks for adverse health outcomes?

Data. We used the cohort identified in Question-1 (participants who had valid answers to all the SDoH questions). However, examination of the SDoH questions revealed that some of them (e.g., cannot afford dental care, cannot afford prescriptions) had a finer level of granularity compared to others (e.g., single household). As the questions with a finer level of granularity tend to be more strongly co-related to each other, when compared with other questions, they also tend to cluster together more strongly due to their granularity compared to other questions, making interpretation of the subtypes difficult. In contrast, as the SDoH subdomains had a more uniform granularity, and at the right level of abstraction for purposes of triaging those at high risk to the proper services in the clinic, we used them to identify the SDoH subtypes.

Analytical Model. To identify SDoH subtypes, their associations with outcomes and covariates, and their future translation into precision medicine, we used a three-part analytical framework called Heterogenization, Integration, and Translation (HIT). As shown in Fig. 2, the heterogenization step was used to identify the subtypes through the use of bipartite modularity maximization46-48 (see Appendix A for more details), the integration step was used to associate multiple datatypes with the subtype model,65 and the translation step was used to qualitatively interpret the subtypes,65 with the goal of developing a future decision-support system to translate the subtypes into clinical practice. The following describes the specific methods used in each of the HIT steps:

1. Heterogenization

Fig. 2. The three steps of the HIT framework to analyze SDoH. (1) Heterogenization of the data to identify subtypes. (2) Integration of multiple datatypes such as from EHRs (e.g., depression), genomic data (e.g., mutated genes), and state (e.g., to determine Medicaid expansion) to determine risk and enrichment of each subtype, and (3) Translation of subtypes through interpretation and predictive modeling, with the goal of designing of clinical decision-support systems for precision medicine.
1. **Heterogenization: Identification of Subtypes.** As there were many participants that did not have valid answers to the SDoH questions, dropping them resulted in differences in the proportion of demographic variables compared with the full All of Us cohort. The data therefore needed to be adjusted to better reflect the overall All of Us participants. To adjust the demographic distribution of the cohort to match the full All of Us cohort, we calculated the inverse probability weights (IPW)\(^{66,67}\) for each participant in the cohort. IPW calculates weights to proportionally boost the values of participants that are underrepresented in the cohort, with respect to a comparison such as the full All of Us cohort, using the method similar to an earlier study on the All of Us cohort\(^{68}\) (see Appendix E). Next, we multiplied the IPW generated weights with the original binary values for each participant in our cohort, and used min-max to range-normalize those weights within each SDoH subdomain. Finally, to test the replicability of the SDoH subdomain biclustering, we randomly divided the dataset into a training and a replication dataset.

We identified subtypes in the training dataset, and tested the degree to which the SDoH subdomain co-occurrences replicated in the test dataset using the following steps: (1) modelled participants and SDoH subdomains as a weighted bipartite network (see Step-1 in Fig. 2) where nodes were either participants (circles), or SDoH subdomains (triangles), and the associations between participant-SDoH subdomain pairs were weighted edges (lines) generated from IPW; (2) used a bipartite modularity maximization algorithm,\(^{46-48}\) (which takes edge weights into consideration) to identify the number of biclusters, their members, and measure the degree of biclusteredness through bicluster modularity (Q, defined as the fraction of edges falling within a cluster, minus the expected fraction of such edges in a network of the same size with randomly assigned edges); (3) measured the significance of Q by comparing it to a distribution of the same quantity generated from 1000 random permutations of the network, while preserving the network size (number of nodes), and the distribution of weighted edges for each participant; (4) used the Rand Index (RI) to measure the degree to which SDoH occurred and did not co-occur in the same cluster in the training and test datasets.; and (5) measured the significance of RI by comparing it to the mean of a distribution of the same quantity generated by randomly permuting the training and replication datasets 1000 times, while preserving the size of the networks.

2. **Integration: Risk and Enrichment of Subtypes.** We used logistic regression to measure the odds ratio (OR) for each subtype compared pairwise to each of the other subtypes, for the three outcomes (Depression, Delayed Medical Care, and ER Visits in Last Year), and for living in a state with Medicaid expansion. To adjust for the difference in demographics due to the missingness, we used weights generated from IPW for each participant, and the comparisons were adjusted for demographics (age, sex, race) and corrected for multiple testing within each outcome using FDR. As 1688 (13.1%) participants did not have 3-digit zip code information, we used IPW and the comparisons were adjusted for potential sample selection bias.

3. **Translation: Interpretation of Subtypes.** The subtype interpretation was done using the following steps: (a) used the Fruchterman-Reingold\(^{49}\) and ExplodeLayout\(^{60,61}\) algorithms to visualize the bipartite network along with the risk for each of the outcomes; and (b) asked the domain experts to independently label the subtypes and infer their mechanisms with potential strategies to address modifiable SDoH with the goal of reducing the risk for each of the outcomes (Depression, Delayed Medical Care, and ER Visits in Last Year), and then collaboratively come to a consensus.

### Results

#### Question-1: What is the range and response to survey questions related to SDoH?

**Identification and Coding of SDoH Questions and Subdomains.** The three domain experts identified 110 questions from 4 surveys (The Basics, Overall Health, Healthcare Access & Utilization, and SDoH). Of these, 110 were abbreviated, and 48 were negatively-worded and coded (see Appendix C). The 110 SDoH questions were further categorized into 19 SDoH subdomains (one of these was Delayed Medical Care that was used as an outcome).

**Response to SDoH Questions and Subdomains.** As shown in Figure 3a, the number of valid responses for each of the 110 SDoH questions was largely dictated by the surveys in which they were solicited. SDoH from 2 surveys (The Basics, Overall Health) had the most valid responses (mean=349434, SD=23556), followed by Healthcare Access & Utilization (mean=149898, SD=6146), and finally the SDoH survey (mean=55960, SD=1083). This pattern of responses match how answers to each of the surveys were solicited: at enrollment,
all participants are required to do The Basics, and Overall Health surveys, and then on a rolling basis the other surveys responses are solicited. The SDoH survey is the latest survey that was solicited, which explained their lowest number of responses. As shown in Fig. 3b, this pattern of missingness held for the responses at the SDoH subdomain level, which was not unexpected as the SDoH subdomains were aggregations of the SDoH questions. However, as shown in Fig. 3A and 3B by the uneven number of valid responses within each survey block, there were several SDoH questions that had invalid responses (“skip” or “chose not to answer”) at both levels of granularity: The Basics: 6%; Health Access & Utilization: 6.1%; Overall Health: 4.39%; and SDoH: 2.61%. Furthermore, the proportion of valid to invalid responses between them as significantly different for both the SDoH questions ($\chi^2 (2, N=365237)=57.489, P<.001$), and for SDoH subdomains ($\chi^2 (2, N=372063)=75.637, P<.001$).

Range of SDoH Questions and Subdomains. As shown by the colored bars in Figure 3, the surveys spanned the full range of the five SDoH Healthy People 2030 domains. The SDoH questions in The Basics and Overall Health surveys were predominantly related to economic stability (blue) and social and community context (purple), those in Healthcare Access & Utilization survey were all related to that topic (green), whereas those from the SDoH survey were a mix of all four domains. Overall, the five Healthy People 2030 domains had strong coverage across the four surveys: Social and Community Context=38; Neighborhood and Built Environment=19; Economic Stability=10; Education Access and Quality=2; Health care Access and Quality=42. This characterization suggests that while the SDoH in All of Us have broad domain coverage across the surveys, analysis of them requires access to all four surveys, which have different levels of completion and valid responses.

Cohort with Maximized Valid Responses. Given the large degree of missingness in 2 of the 4 surveys, we could not use multiple imputation to fill in the values, and therefore had to find a subset of participants that had valid responses to all the SDoH questions. An examination revealed that two SDoH questions had <10% responses (English Verbal Frequency=1.67%, and Neighborhood has no recreation spaces=8.4%), accounting for the largest loss in cohort size with valid responses. These questions were therefore dropped from further analysis. Furthermore, one question required a branched response (Living Situation branching to Did not Live in a House) which were merged. Finally, as we used Delayed Medical Care as an outcome, 9 remaining questions related to that topic were removed, resulting in a total of 98 SDoH questions. As shown in Fig. 4, a Venn diagram of the overlap among the valid responses across the surveys revealed that 12,913 participants had valid responses to all 98 SDoH questions.
Co-occurrence of the Number of SDoH across Responders. As shown in Fig. 5, participants had a median of 15 SDoH question co-occurrences and a median of 9 SDoH subdomains co-occurrences. Furthermore, participants of color or racial/ethnic minorities, who had valid responses to the 110 SDoH questions, had a significantly higher median number of co-occurring SDoH compared to the equivalent White population (median participants of color or racial/ethnic minorities=20, median White=14, \(P < .001\)). These results show the high co-occurrences of SDoH at both levels of granularity, with a significant difference in median co-occurrences between the White and the participants of color or racial/ethnic minority populations, with valid responses.

Participant Demographics with Valid Responses to SDoH Questions. As the cohort size dropped to 3.5%, we analyzed how that impacted the demographic distribution compared with the overall All of Us cohort. As shown in Table 2, there were statistically significant differences in race (\(\chi^2(5, N=372,397)=2073.1, P < .001\)), and ethnicity (\(\chi^2(9, N=372,397)=6292.2, P < .001\)) between the two cohorts, after multiple testing correction, with a higher proportion of White participants having valid answers compared to participants of color or racial/ethnic minorities. Furthermore, there was a statistically significant difference in age between the participants who had valid answers, versus those that did not (\(H(1)=148.08, P < .001\)). These results show the demographic differences between the cohort with complete and valid answers to the SDoH questions, in comparison to the full All of Us cohort.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>All AoU Participants: 372,397 (100%)</th>
<th>All AoU Participants with valid(^a) SDoH answers: 12,913 (3.5%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Race</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>201149 (54.01%)</td>
<td>11279 (87.35%)</td>
</tr>
<tr>
<td>Black or African American</td>
<td>73383 (19.71%)</td>
<td>482 (3.73%)</td>
</tr>
<tr>
<td>Asian</td>
<td>12459 (3.35%)</td>
<td>324 (2.51%)</td>
</tr>
<tr>
<td>Other or &gt;1 population</td>
<td>26890 (7.22%)</td>
<td>343 (2.66%)</td>
</tr>
<tr>
<td>None Indicated</td>
<td>58516 (15.71%)</td>
<td>485 (3.76%)</td>
</tr>
<tr>
<td><strong>Ethnicity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Hispanic or Latino</td>
<td>288227 (77.4%)</td>
<td>12095 (93.67%)</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>66704 (17.91%)</td>
<td>751 (5.82%)</td>
</tr>
<tr>
<td><strong>Additional Options</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>138831 (37.28%)</td>
<td>4674 (36.09%)</td>
</tr>
<tr>
<td><strong>Sex at birth</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>222495 (59.75%)</td>
<td>8236 (63.6%)</td>
</tr>
<tr>
<td>Male</td>
<td>138831 (37.28%)</td>
<td>4674 (36.09%)</td>
</tr>
<tr>
<td>Intersex</td>
<td>80 (0.02%)</td>
<td>20 (0.15%)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>220833 (59.3%)</td>
<td>8113 (62.82%)</td>
</tr>
<tr>
<td>Male</td>
<td>138140 (37.09%)</td>
<td>4642 (35.95%)</td>
</tr>
<tr>
<td>Non Binary</td>
<td>920 (0.25%)</td>
<td>60 (0.46%)</td>
</tr>
<tr>
<td>Transgender</td>
<td>464 (0.12%)</td>
<td>20 (0.15%)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median=56 (19-122(^a))</td>
<td>Median=58 (19-93)</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Participants that completed all questions, and did not skip, or choose not to answer a question; \(^b\)Age 122 = a participant chose the least birth year (1900).
Question-2: How do SDoH subdomains co-occur to form subtypes, and what are their risk for adverse health outcomes?

The cohort used to identify the subtypes consisted of the 12,913 participants, of which 12,886 that had valid IPW weights. These were split randomly into the training and replication datasets with complete data for 18 SDoH subdomains (identified in Question-1), in addition to the three outcomes (depression, delayed medical care and ER visits in last year), and covariates (demographics).

1. Heterogenization: Identification of Subtypes

The weighted (using IPW generated weights) bipartite network analysis of the training dataset (n=6492) and the 19 SDoH subdomains revealed 4 biclusters with statistically significant bicluster modularity (Q=0.13, random-Q=0.11, z=7.5, P<0.001) and the SDoH subdomains significantly replicated in the replication dataset (Real-RI=0.88, Random-RI=0.62, P<0.001). Across all three outcomes, Cluster-1 had a significantly higher OR compared to Cluster-4. The labels in bold text represent the consensus interpretation of the clusters by three domain experts.

![Fig. 6. Four biclusters in the training dataset consisting of subgroups of participants (n=6492), and their most frequently co-occurring SDoH subdomains (d=18). The biclustering was significant (Q=0.13, random-Q=0.11, z=7.5, P<0.001) and the SDoH subdomains significantly replicated in the replication dataset (Real-RI=0.88, Random-RI=0.62, P<0.001). Across all three outcomes, Cluster-1 had a significantly higher OR compared to Cluster-4. The labels in bold text represent the consensus interpretation of the clusters by three domain experts.](image-url)

2. Integration: Risk and Enrichment of Subtypes

Table 3 shows the association of each subtype to the three outcomes. As shown by the dark orange row, Cluster-1 (housing, food security, income, literacy, employment, education attainment) had a significantly higher OR for each of the three outcomes compared to Cluster-4 (provider characteristics, neighborhood characteristics, healthcare coverage, demographics, supportive relationships, language). Furthermore, within the Depression outcome, each of the clusters had a significantly higher OR compared to one other cluster forming a ranking of risk among all the four clusters (1>3>2>4). In contrast, Delayed Medical Care had two other significant
associations (2>1, 3>4), with ER Visit in the Last Year having only the one significant pair-wise association that fit into the overall trend.

As shown in Table 4, this trend continued in the enrichment analysis of association with living in a state with No Medicaid Expansion. As shown, Cluster-1 had a significantly higher OR compared to Cluster-4, in addition to the other clusters. The overall results suggests that Cluster-1 and Cluster-4 form “book ends” representing the high and low ends of risk among the clusters.

3. Translation: Interpretation of SDoH Subtypes and Design of Potential Interventions

The three domain experts examined the co-occurrences of SDoH subdomains within each bicluster shown in the network visualization (Fig. 6), and integrated them with the quantitative ORs in Table 3 and 4. The consistent “book ends” result where Cluster-1 had significantly higher ORs compared with Cluster-4 across all four variables was of strong interest, and interpreted as follows: (1) **Cluster-1** was labeled **Socioeconomic Barriers** as it contained multiple high risk SDoH. These co-occurring SDoH could have resulted from cascades over time such as low education leading to unstable employment, which in turn led to low income, and food and housing insecurity. Such cascading factors could be perceived by participants to be unmodifiable, leading to chronic stress and depression. Furthermore, the strong association of this subtype with the outcomes Delayed Medical Care and ER Visits in the Last Year, and that participants in this subtype were more likely to be from a US state with No Medicaid Expansion, began to provide a more comprehensive understanding of this high-risk SDoH subtype. For example, individuals that match this subtype profile could be screened for depression early for and simultaneously triaged to social services prioritized to address specific SDoH triggers such as food security, housing, and income if feasible. (2) **Cluster-4** was labeled **Sociocultural Barriers** as it contained a combination of SDoH related to neighborhood and supportive relations, in addition to cultural barriers related to language, provider interactions, and language. In contrast to the more devastating socioeconomic barriers in

<table>
<thead>
<tr>
<th>Cluster Comparison</th>
<th>Outcomes</th>
<th>Depression OR=0.79, CI=0.64-0.97, P-corr=0.022 &lt;.05</th>
<th>Delayed Medical Care OR=3.5, CI=3.4-1.1, P-corr&lt;0.0038 &lt;.001</th>
<th>ER Visit in the Last Year OR=1.8, CI=1.4-2.3, P-corr=0.0016 &lt;.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster A vs. Cluster B</td>
<td></td>
<td>Cluster A vs. Cluster B</td>
<td>Cluster A vs. Cluster B</td>
<td>Cluster A vs. Cluster B</td>
</tr>
<tr>
<td>1 vs 2</td>
<td>OR=1.7, CI=1.5-2.1, P-corr=2.5e-10 &lt;.001</td>
<td>OR=0.78, CI=0.67-0.92, P-corr=0.0038 &lt;.001</td>
<td>OR=1.2, CI=0.91-1.6, P-corr=0.24</td>
<td></td>
</tr>
<tr>
<td>1 vs 3</td>
<td>OR=1.3, CI=1.1-1.6, P-corr=0.019 &lt;.05</td>
<td>OR=0.88, CI=0.72-1.1, P-corr=0.23</td>
<td>OR=1.4, CI=0.96-1.9, P-corr=0.13</td>
<td></td>
</tr>
<tr>
<td>1 vs 4</td>
<td>OR=4.2, CI=3.5-5.1, P-corr=3.5e-52 &lt;.001</td>
<td>OR=3.5, CI=3.4-1.1, P-corr=1.8e-53 &lt;.001</td>
<td>OR=1.8, CI=1.4-2.3, P-corr=0.0016 &lt;.001</td>
<td></td>
</tr>
<tr>
<td>2 vs 3</td>
<td>OR=0.79, CI=0.64-0.97, P-corr=0.022 &lt;.05</td>
<td>OR=1.2, CI=0.98-1.4, P-corr=0.094</td>
<td>OR=1.0, CI=0.75-1.5, P-corr=0.8</td>
<td></td>
</tr>
<tr>
<td>2 vs 4</td>
<td>OR=2.3, CI=1.9-2.7, P-corr=3.2e-21 &lt;.001</td>
<td>OR=4.3, CI=3.7-5, P-corr=3.1e-35 &lt;.001</td>
<td>OR=1.3, CI=1.1-1.7, P-corr=0.12</td>
<td></td>
</tr>
<tr>
<td>3 vs 4</td>
<td>OR=2.9, CI=2.3-3.5, P-corr=1.5e-21 &lt;.001</td>
<td>OR=3.6, CI=3.4-4.4, P-corr=1.6e-38 &lt;.001</td>
<td>OR=1.4, CI=0.95-1.9, P-corr=0.13</td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Across all three outcomes, Cluster-1 had a significantly higher risk compared to Cluster-4 (dark orange row). The Depression outcome had a distinct ranking of risks (light orange), whereas the other two outcomes had a subset of them.

<table>
<thead>
<tr>
<th>Cluster Comparison</th>
<th>Enrichment</th>
<th>No Medicaid Expansion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster A vs. Cluster B</td>
<td></td>
<td>Cluster A vs. Cluster B</td>
</tr>
<tr>
<td>1 vs 2</td>
<td>OR=1.5, CI=1.3-1.8, P-corr=1.7e-5 &lt;.001</td>
<td>OR=0.5, CI=0.3-0.8, P-corr=0.0085 &lt;.001</td>
</tr>
<tr>
<td>1 vs 3</td>
<td>OR=1.3, CI=1-1.6, P-corr=0.048 &lt;.05</td>
<td>OR=1.3, CI=1-1.6, P-corr=0.057 &lt;.01</td>
</tr>
<tr>
<td>1 vs 4</td>
<td>OR=0.9, CI=0.8-1.3, P-corr=0.97</td>
<td>OR=1.4, CI=0.8-2-1.2, P-corr=0.97</td>
</tr>
<tr>
<td>2 vs 3</td>
<td>OR=0.99, CI=0.8-1.2, P-corr=0.97</td>
<td>OR=1.4, CI=0.8-2-1.2, P-corr=0.97</td>
</tr>
<tr>
<td>2 vs 4</td>
<td>OR=0.99, CI=0.8-1.2, P-corr=0.97</td>
<td>OR=1.4, CI=0.8-2-1.2, P-corr=0.97</td>
</tr>
</tbody>
</table>

Table 4. Cluster-1 had a significantly higher OR compared to Cluster-4 (dark orange) for no Medicaid expansion, in addition to Cluster-2 and Cluster-3 (light orange).
While Cluster-1 and Cluster-4 formed the “book ends” of risk across the three outcomes, Cluster-2 was flagged as critical and labeled Lived Experience Barriers. The SDoH in this cluster included discrimination in everyday life, medical settings, in addition to problems in interactions with providers and affording medical care. These frequently co-occurring SDoH could explain why this subtype had a significantly higher OR for Delayed Medical Care compared to Cluster-1. Finally, Cluster-3 was labeled Social Context Barriers as the SDoH related to neighborhood cohesion and interactions with others. While not as critical as Cluster-1 and Cluster-2, this cluster still had significantly higher OR for depression compared to Cluster-4.

The clusters were therefore meaningful, and additionally did not have a one-to-one mapping to the 5 SDoH domains defined by Healthy People 2030. As shown in Fig. 7, these data-driven clusters have a complex relationship with the SDoH domains and subdomains. While one subtype belonged to a single domain (subtype Social Context belonged to the domain Social and Community Context), three of the four subtypes belonged to two or more domains (e.g., the subtype Socioeconomic Barriers belonged to the domains Economic Stability, and Education Access and Quality). Such interdomain relationships reflect how SDoH co-occur in the real world reflecting the complex cross-domain interactions described in the Dahlgren-Whitehead model (Fig. 1). These relationships could be useful for refining conceptual models to explain the complex association between SDoH and adverse health outcomes, and to build more accurate SDoH models for predicting adverse health outcomes.

F. Discussion

The mechanisms through which SDoH translate into adverse health outcomes is complex consisting of many interacting factors and feedback loops among individual and environmental/contextual factors. While this phenomenon has been studied for more than three decades, the limited range of data types, limited representation of populations that have been socially marginalized, and limited access to individual-level data at scale due to privacy laws have been critical hurdles for researchers. Recognizing that All of Us has a well-articulated plan and resources to overcome these limitations, but is still in a rapidly evolving stage, we conducted a systematic characterization of more than a hundred SDoH available in All of Us, and used them to identify SDoH subtypes with the future goal of designing precision medicine interventions. This attempt led to the following opportunities and challenges related to data, methods, and theory.

Data: Missingness and Granularity

All of Us data contained 110 SDoH across 4 surveys, and 93 SDoH-related SNOMED codes in the EMRs. While these provided a comprehensive coverage of SDoH with respect to domains and subdomains identified by Healthy People 2030, our analysis uncovered the following patterns of missingness and SDoH granularity.

Missingness. The analysis revealed three types of missingness: (1) Rollout Missingness: This type of missingness was largely dictated by how the surveys were rolled out to participants. As all participants at enrollment are required to do The Basics, and Overall Health surveys, they had the highest responses, followed by the later solicited survey Healthcare Access & Utilization, and the SDoH survey rolled out more recently in 2022. This order of rollout was the main source of missingness resulting in a precipitous reduction in cohort size for those that had answers to all the SDoH questions. (2) Valid Answer Missingness. As participants can choose not to answer any survey questions, the data contained “PMIs” related to “skip” and “choose not to answer”. These accounted for a much smaller reduction in cohort size for complete data. (3) Low Usage Missingness. Although there were 259 SDoH SNOMED codes, only 93 (3.3%) had such information for >20 participants that are allowed to be reported. This could be because most clinicians currently do not screen for SDoH, as it is typically done by the social worker. Furthermore, we also attempted to use 3-digit zip codes to determine which subtypes had a significant association to living in a state that did not offer Medicaid expansion, and a mutation on the gene. However, 13.1% (1688) of the participants did not have zip code information (adjusted by using IPW), and 3248 (25.2%) did not have genetic information (which we are currently in the process of analyzing).

Together, the above three types of missingness impacted the size of the resulting cohort that had valid answers, in the following two ways: (1) a drastic reduction in cohort size by 93.5%. However, because of the size of the overall cohort (n=372,397), we were still left with a large subcohort (n=12,886), which to the best of our knowledge is the largest subset of individuals to be analyzed for such a wide range of SDoH; and (2) significant differences in the proportion of race, ethnicity, and age in the above cohort when compared to the overall All of Us population. Specifically, the cohort with valid answers had significantly more White, or non-Hispanic, or older participants, when compared to the overall cohort. This could potentially be because once a participant has been
enrolled, there is a 90-day delay in sending subsequent solicitations to complete surveys, a policy that is currently being re-assessed due to its impact on missingness. We therefore had to correct this imbalance in demographics by using IPW for identifying subtypes that were representative of the overall All of Us cohort.

Granularity. Because our goal was to use machine learning methods to identify SDoH subtypes, we encountered uneven granularity in the SDoH questions. Some questions were fine-grained and highly correlated and therefore would cluster more strongly because of the nature of the granularity of the questions, not because of the SDoH mechanisms. To address this uneven granularity, and to make the results more interpretable, we used SDoH subdomains which had a coarser but more consistent level of granularity. We chose this approach because SDoH subdomains had already been defined, were understood by domain experts, and appeared to be at the right level of abstraction useful for clinical applications such as triaging a patient to the right social service. However, because the use of coarse-grained variables loses information, future research could explore aggregating only those SDoH questions that are highly correlated, and preserving the rest at the finer level of granularity.

Method: Scalability, Generalizability, and Extensibility

We designed the HIT analytical framework to be scalable enabling its use for the growing size of the cohort in All of Us, to be generalizable across cohorts and conditions, and to be extensible for including additional methods as needed in the future. Testing the HIT framework on the All of Us data provided insights for the strengths and limitations of the framework, and for the All of Us workbench where the analysis was conducted.

Scalability. We used three types of code to conduct the analysis for both research questions. (1) Automatically generated code to extract the cohort, produced by All of Us once a cohort was selected using the point and click interface. This code was adequately scalable and generalizable and so will not be discussed further. (2) Customized code to extract specific parts of the data. For example, the analysis of co-occurrences required customized code in R to plot the diagrams in Fig. 3. As expected, these tasks required strong programming skills, but fortunately we did not encounter any coding or execution problems using the R or Python programming languages. However, there were significant server issues which hampered our analysis. Although the workbench instructions state that code running on the workbench for more than 2 weeks would be terminated and all intermediate results deleted, we frequently encountered our work disappearing at shorter intervals. These disruptions resulted in a higher consumption of the free server time credits, and fewer analyses that we could conduct due to the computation time. (3) Machine learning code we had previously developed and disseminated on CRAN69-71 to conduct the bipartite network analysis and the significance testing, and to visualize the network. As this code was designed to be generalizable and scalable, we did not encounter any issues in the execution of our code (besides the same server issues mentioned above). Finally, the visualization of our networks worked as expected, and we used them to help interpret the patterns in the data.

Generalizability. Our code for the first two steps of the HIT framework is in Jupyter notebooks and have been used to analyze other cohorts that were filtered for age and prior conditions. For example, we extracted a cohort (n=4090) of participants with diabetes aged >=65 with complete data on 18 SDoH variables selected through consensus by 2 experienced health services researchers, and guided by Andersen’s behavioral model. The analysis72,73 revealed 7 SDoH subtypes with statistically significant modularity compared with 100 random permutations of the data (All of Us=.51, Random Mean=.38, z=20, P<.001), and which were not only clinically meaningful, but also significant in different degrees for the outcome. Our subsequent attempt at increasing the number of SDoH variables from 18 to 110 for participants with diabetes that had valid answers, led to an extremely small cohort size (n=926) (see Appendix D) due to the missingness that we described above. While this reduction resulted in our current strategy of analyzing all participants regardless of condition or age, these experiments demonstrate that our approach is generalizable to other subsets of the data.

Extensibility. The HIT model is designed to be extensible to include other methods. For example, the model could use other biclustering (e.g., Non-negative Matrix Factorization4) and causal modeling methods, and use different types of classification (e.g., deep learning36), and prediction methods (e.g., subgroup-specific modeling 36) to build the decision-support system in the Translational Step (Fig. 2). Furthermore, the model can integrate a wide range of data types to enable analysis of how each subtype is associated with them, resulting in a layered interpretation of the SDoH subtypes as we have demonstrated. Furthermore, the integration of different subtypes required a diverse team consisting of experts in machine learning, biostatistics, programming, clinical care, health services research, and gerontology to enable a 360 analysis and interpretation of the subtypes. The use
of the workbench to share results operationalized team-centered informatics designed to facilitate multidisciplinary translational teams to work more effectively across disciplinary boundaries, with the goal of analyzing and designing interventions for precision medicine.

**Theory: Model Building, and Translational Implications**

The identification of SDoH subtypes has strong implications for model building in addition to translational applications. As shown in Fig. 7, while the current classification of five SDoH domains has a hierarchical relationship with the SDoH subdomains, the data-driven clusters have a more complex association with the same SDoH subdomains. This reflects the complexity of how SDoH occur in the real-world.

Future models should develop predictive models using both the knowledge-based and cluster-based results, to determine which is more accurate for predicting adverse outcomes. Because the subtypes were clinically interpretable, they could be used to build classification and predictive models, and used with an interface to develop a clinical decision-support system that help to triage patients to critical services. For example, the St. Vincent House (https://www.stvhope.org/) in Galveston, Texas provides several services to address SDoH including free walk-in clinical care, nurse practitioner with small co-pay requested, English and Spanish-speaking free mental health counseling, free dental health clinic, utility and rental assistance, case management, financial literacy, expanded food pantry, weekly free home delivery of pantry groceries, snack pack for people experiencing homelessness, free transportation for doctor's appointment, immigration legal services, and spiritual counseling. Given the availability of this wide range of services in many communities across the US, a decision-support system could help to classify an individual based on their SDoH profile into one or more of the subtypes, measure their risk for an adverse health outcome, and proactively recommend one or more of such local services (Fig. 2, Step-3).

**Notebooks for All of Us Community Use**

Because the missingness of SDoH should reduce in the near future, their characterization and subtyping will need to be repeated and verified for different cohorts. Therefore, we have made the following two sets of code available for general use by All of Us researchers:

1. **SDoH Valid Answer Tracker.** This set of notebooks generate four plots which can be used by other researchers on All of Us to characterize any cohort: (1) valid responses plot to show how many participants have data with valid responses, and colored by SDoH domains; (2) Venn diagram showing how many participants have valid responses for all questions within each survey; (3) frequency distribution plot showing co-occurrence of SDoH across the selected cohort. This set of tools should enable researchers to characterize SDoH across different cohorts, to help determine methods that are appropriate to adjust for missingness in those cohorts.

2. **SDoH Subtyper.** This set of notebooks can be used to conduct the following analyses: (1) bicluster modularity of a cohort with the 18 SDoH subdomains to identify the number and members of biclusters, and the measure Q representing the quality of the biclustering; (2) visualization of the bipartite network; and (3) significance of the network with respect to null models.

**Limitations**

This study has two main limitations. The first emerges from the temporary limitations of the large amount of missingness in the survey data, precluding the use of imputation methods. We could therefore use only complete data, which led to a large drop in cohort size, and which also introduced a bias in the demographics requiring a rebalancing through IPW. While such rebalancing is typically done for large datasets, the IPW method requires judgement to decide which variables to include in the model, and therefore could have introduced additional unknown biases. Therefore, the model should be refined to determine which variables to include in the regression models that estimate the IPWs. However, because the clustering was similar between the unweighted and IPW weighted networks, we believe that the clusters are stable, meaningful, and represent the full All of Us cohort. This limitation is expected to be addressed as All of Us has recently removed the requirement of waiting for 90 days before a subsequent survey is given to an enrollee in the program, potentially reducing the degree of missingness. The second limitation is due to the high computational cost of empirically determining the significance of the biclustering. As such analysis is computationally expensive and time-consuming, it limited the experiments we could do to test different cohorts and models. We therefore look forward to the All of Us workbench providing the ability to run batch processes more efficiently, and which will be uninterrupted for
extended periods of time (exceeding the current time window), which together could help alleviate this computational hurdle in the future.

G. Conclusion

How SDoH impact health is a complex phenomenon involving many interconnected factors which have yet to be fully elucidated. While this phenomenon has been studied for more than 30 years, the analyses have been hampered by the lack of large cohorts representing diverse populations with multiple datatypes, and with easy access by researchers. *All of Us* provides an unprecedented opportunity to directly address these limitations with the goal of doing justice to early conceptual models such as the social gradient and the Dahlgren-Whitehead model, both of which drew international attention to the complex ways in which individual and contextual SDoH factors impact health. The *All of Us* dataset is also timely because of the visceral health disparities that were revealed during the pandemic, which pushed the abstract phenomenon of SDoH into the public and policy realms. However, because *All of Us* is still rapidly evolving to meet its target of one million participants or more, we conducted a systematic characterization of SDoH variables in *All of Us*, and used the results to guide the analysis of SDoH subtypes, with the future aim of designing precision medicine interventions.

Our first goal of characterizing the data revealed the nature of the missingness in SDoH, and the uneven granularity in the SDoH questions. Both these results led us to select the IPW method to address the missingness, and analysis of subtypes at the SDoH subdomain level of granularity. Our second goal of identifying SDoH subtypes led not only to statistically significant biclusters, but also to their statistically significant replication, and meaningful domain interpretations. These results set the stage for further investigations to build and evaluate prediction models, and decision-support systems. The results also led to the design, use, and dissemination of general-purpose tools currently available on *All of Us* for other researchers, which will be useful to reanalyze the *All of Us* data as it grows over the next few years. These collaborative advances should position *All of Us* to revolutionize research for analyzing complex phenomena such as how SDoH impact health and beyond, with the goal of enabling a more equitable future that all of us deserve.

H. References


63. Decker A, Weaver R. Health and Social Determinants Associated With Delay of Health Care Among Rural Older Adults. Innovation in Aging. 2021;5(Supplement_1):210-211.


Appendix A: Description of Bipartite Network Analysis

A network consists of nodes and edges; nodes represent one or more types of entities (e.g., participants or SDoH), and edges between the nodes represent a specific relationship between the entities. Figure 1A shows a unipartite network where nodes are the same type (typically used to analyze co-occurrence of comorbidities). In contrast, Figure 1B shows a bipartite network where nodes are of two types, and edges exist only between different types such as between participants (circles) and SDoH (triangles). This quantitative and visual representation, which integrates patients and their SDoH in a single representation, enables domain experts to infer the mechanisms in each patient subgroup, and design targeted interventions, a cornerstone of precision medicine.

Fig. 1. The distinction between a unipartite network (A), a bipartite network (B), and how the latter can be used to identify biclusters of participants and SDoH (C).
Appendix B: SNOMED Codes Related to SDoH, and their Use in the Electronic Health Records of Participants in All of Us

![SNOMED CodesRelated to SDoH](image-url)
Appendix C: SDoH questions extracted from 4 All of Us surveys, abbreviated, negatively labeled and reverse coded, categorized into the Healthy People 2030 domains, and further categorized into subdomains.

<table>
<thead>
<tr>
<th>Survey Name</th>
<th>Question/Field</th>
<th>Abbreviated and Negatively Labeled</th>
<th>SDoH Domain (healthy.gov)</th>
<th>Sub Domain (Domain Knowledge)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social Determinants of Health</td>
<td>People around here are willing to help their neighbors</td>
<td>Neighborhood people unwilling to help</td>
<td>Social &amp; Community Context</td>
<td>Neighborhood cohesion</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>People in my neighborhood generally get along with each other</td>
<td>Neighborhood people do not get along</td>
<td>Social &amp; Community Context</td>
<td>Neighborhood cohesion</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>People in my neighborhood can be trusted</td>
<td>Neighborhood people cannot be trusted</td>
<td>Social &amp; Community Context</td>
<td>Neighborhood cohesion</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Neighbors people do not share values</td>
<td>Neighborhood people do not share values</td>
<td>Social &amp; Community Context</td>
<td>Neighborhood cohesion</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>People who are always having trouble with my neighbors</td>
<td>Neighborhood people have trouble</td>
<td>Social &amp; Community Context</td>
<td>Neighborhood cohesion</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>In my neighborhood, people watch out for each other</td>
<td>Neighborhood people do not watch out</td>
<td>Social &amp; Community Context</td>
<td>Neighborhood cohesion</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Someone to help you if you were confined to bed</td>
<td>No one to help out of bed</td>
<td>Social &amp; Community Context</td>
<td>Supportive Relationships</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Someone to take you to the doctor if you need it</td>
<td>No help for doctor visit</td>
<td>Social &amp; Community Context</td>
<td>Supportive Relationships</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Someone to prepare your meals if you were unable to do it yourself</td>
<td>No one to help with meal prep</td>
<td>Social &amp; Community Context</td>
<td>Supportive Relationships</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Someone to help with daily chores if you were sick</td>
<td>No one to help with daily chores if you were sick</td>
<td>Social &amp; Community Context</td>
<td>Supportive Relationships</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Someone to have a good time with</td>
<td>No one to have a good time with</td>
<td>Social &amp; Community Context</td>
<td>Supportive Relationships</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Someone to turn to for suggestions about how to deal with a personal problem</td>
<td>No one to suggest problem solutions</td>
<td>Social &amp; Community Context</td>
<td>Supportive Relationships</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Someone who understands your problems</td>
<td>No one who understands your problems</td>
<td>Social &amp; Community Context</td>
<td>Supportive Relationships</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Someone to make you feel wanted</td>
<td>No one to make you feel wanted</td>
<td>Social &amp; Community Context</td>
<td>Supportive Relationships</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Lack companionship</td>
<td>Lack companionship</td>
<td>Social &amp; Community Context</td>
<td>Relationships with others</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>You are treated with less courtesy than other people are</td>
<td>Others think you are less courteous</td>
<td>Social &amp; Community Context</td>
<td>Discrimination in everyday life</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Day to day less courteous frequency</td>
<td>Others think you are less respectful</td>
<td>Social &amp; Community Context</td>
<td>Discrimination in everyday life</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>You receive poorer service than other people at restaurants or stores</td>
<td>Others think you are not treated with respect</td>
<td>Social &amp; Community Context</td>
<td>Discrimination in everyday life</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>People act as if they think you are not smart</td>
<td>Others think you are not smart</td>
<td>Social &amp; Community Context</td>
<td>Discrimination in everyday life</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>People act as if they are afraid of you</td>
<td>Others are afraid of you</td>
<td>Social &amp; Community Context</td>
<td>Discrimination in everyday life</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>People act as if they think you are dishonest</td>
<td>Others think you are dishonest</td>
<td>Social &amp; Community Context</td>
<td>Discrimination in everyday life</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>People act as if they’re better than you are</td>
<td>Others think they’re better than you are</td>
<td>Social &amp; Community Context</td>
<td>Discrimination in everyday life</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>You are called names or insulted</td>
<td>Called names or insulted</td>
<td>Social &amp; Community Context</td>
<td>Discrimination in everyday life</td>
</tr>
<tr>
<td>Social Determinants of Health</td>
<td>Threatened or harassed frequently</td>
<td>Threatened or harassed frequently</td>
<td>Social &amp; Community Context</td>
<td>Discrimination in everyday life</td>
</tr>
</tbody>
</table>

The copyright holder for this preprint this version posted January 28, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .
Many shops, stores, markets or other places to buy things I need are within easy walking distance of my neighborhood.

During the past 12 months, was there any time when you needed dental care but didn’t receive it because you couldn’t afford to?

When you go to a doctor’s office or other health care provider, how often do you receive timely advice or information about your health?

During the past 12 months, was there any time when you needed to see a specialist but didn’t because you couldn’t afford to?

When you go to a doctor’s office or other health care provider, how often are you treated with respect by your doctors or health care providers?

Health care access and quality

Has your delayed getting care in the past 12 months because you provide care to an adult who is elderly or disabled?

Health care access and quality

Health care access and utilization

During the past 12 months, were you told by a health care provider or doctor’s office that you had to pay out of pocket for some or all of the procedure?

Health care access and quality

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctors or health care providers to give you information about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you get your prescription medicine from a mail order pharmacy to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you alter the dosage of your prescription medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctors or health care providers to give you information about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you get your prescription medicine from a mail order pharmacy to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you alter the dosage of your prescription medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctors or health care providers to give you information about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you get your prescription medicine from a mail order pharmacy to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you alter the dosage of your prescription medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you get your prescription medicine from a mail order pharmacy to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you alter the dosage of your prescription medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you get your prescription medicine from a mail order pharmacy to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you alter the dosage of your prescription medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you get your prescription medicine from a mail order pharmacy to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you alter the dosage of your prescription medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you get your prescription medicine from a mail order pharmacy to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you alter the dosage of your prescription medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you get your prescription medicine from a mail order pharmacy to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you alter the dosage of your prescription medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you get your prescription medicine from a mail order pharmacy to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you alter the dosage of your prescription medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you delay filling a prescription to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you ask your doctor or health care providers about the cost of your prescription medication?

Health care access and quality

Health care access and utilization

During the past 12 months, did you get your prescription medicine from a mail order pharmacy to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you alter the dosage of your prescription medicine to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you use alternative therapies to save money?

Health care access and quality

Health care access and utilization

During the past 12 months, did you take less medicine to save money?
**Appendix D: Condition-Specific Cohort Extraction**

![Diagram showing cohort extraction process]

**Figure 1.** Inclusion and exclusion criteria for selecting three condition-specific cohorts.
Appendix E: Inverse Probability Weighting

We found significant differences in the demographic characteristics between the total All of Us sample, and the 12,913 participants with valid answers for all SDoH questions. To account for potential sample selection bias, we calculated inverse probability weights (IPW) using the ipwpoint function in the R package ipw. This function uses a logistic regression model to estimate the predicted probability of having valid responses on all SDoH variables based on age, sex, race, ethnicity, being born in the United States, currently employed, having a college degree or higher, health insurance, owning a home, and being married. We stabilized the weights according to the observed probability of being in our cohort.