Development and Validation of an Individual Socioeconomic Deprivation Index (ISDI) in the NIH's *All of Us* Data Network

Nripendra Acharya, BA[1], Karthik Natarajan, PhD[1]
[1]Columbia University Medical Center, Department of Biomedical Informatics, New York, New York

1 Background

A little over 15 years ago, the final report produced by the World Health Organization's Commission on Social Determinants of Health identified health equity as a key criterion for social and economic policies and initiatives[1,2,3], which paralleled the growing awareness within the United States that medical care alone cannot address health disparities without an understanding of the fundamental role of social factors on health[4,5,6]. However, given the multi-domain and dynamic nature of its component factors[7,8], along with the complex pathways in which these factors affect health outcomes[9,10], social determinants of health (SDOH) can be challenging to assess and articulate in a clear and measurable framework. Composite indices that aggregate diverse SDOH factors, such as Area Deprivation Index[11-13], Social Vulnerability Index[14], and Community Deprivation Index[15], are valuable tools in providing a quantifiable framework for assessing, monitoring, and comparing health disparities for a variety of purposes[16,17,18,19,20].

In general, composite SDOH indices can include spatially specific environmental variables like water quality and air-pollution[21-23] as well as broadly aspatial socioeconomic variables like education level and income[11,15]. However, many of the most-widely used socioeconomic deprivation indices in the United States are constrained by their reliance on constrained census-based neighborhood definitions, the modifiable area unit problem (MUAP)[24], and their dependency on stale and aggregated American Community Survey (ACS) responses[24]. Furthermore, these approaches possess limitations in (1) understanding disparities between subcommunities that exist within a given area entity[25], (2) comparing shared SDOH profiles for patients in disparate geographic regions[26], and (3) effectively capturing temporal changes for both individual patients and for subcommunities[27,28].

The NIH's All of Us Research Program (*All of Us*) presents an opportunity to construct a composite individual level socioeconomic index, hereafter referred to as ISDI, using a nation-wide data network that includes wide-ranging SDOH factors collected at the participant level. In this study, we focus on two aims: (1) the development of an individual-level socioeconomic deprivation index, and (2) the initial validation of this index. This validation includes two parts: (1) assessing correlation with an area-approximated index, (2) assessing changes in AI model performance and accuracy in the context of stratified sampling based on ISDI quintiles.

2 Methods

2.1 Data Source

This study used the *All of Us* v7 Curated Data Repository (CDR), Controlled Tier Data. *All of Us* has created seven surveys, of which three are available to participants to complete right upon initial enrollment. Furthermore, recent research has shown that missingness was low in *All of Us* baseline surveys, and only 0.2% of participants skipped all questions in at least one of the baseline surveys.[29] The participant- completed survey responses were then converted into the OMOP data format using the PPI vocabulary within the OMOP vocabulary.

2.2 Ontology Mapping & Index Construction

The initial step in index construction was domain selection, which was done through extraction of identified domains from recent systematic reviews and grey literature on area-based socioeconomic deprivation indices.[30,31] The relevant domains identified were income, education, financial stress, household structure, housing and environment, transportation, insurance, and language barrier. Each domain was then mapped to an OMOP concept that represented a PPI answer from one of the following survey modules: *The Basics, Healthcare Access and Utilization, Social Determinants of Health*, which themselves were originally derived from the CDC's BRFSS questionnaire, the Veterans Health Administration's National Homelessness screening instrument, and the California Health Interview Survey.

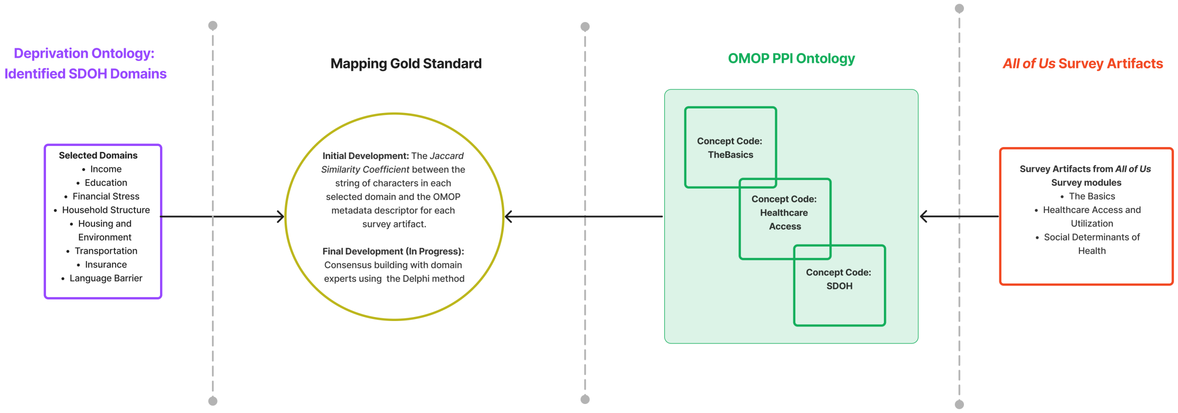

Deprivation Ontology:
Identified SDOH Domains
Selected Domains
• Income
• Education
• Financial Stress
• Household Structure
• Housing and Environment
• Transportation
• Insurance
• Language Barrier
Mapping Gold Standard
Initial Development: The Jaccard Similarity Coefficient between the string of characters in each selected domain and the OMOP metadata descriptor for each survey artifact.
Final Development (In Progress): Consensus building with domain experts using the Delphi method
OMOP PPI Ontology
Concept Code: TheBasics
Concept Code: Healthcare Access
Concept Code: SDOH
All of Us Survey Artifacts
Survey Artifacts from All of Us Survey modules
• The Basics
• Healthcare Access and Utilization
• Social Determinants of Health

Figure 1. Ontology mapping framework between relevant SDOH domains (the Deprivation Ontology) and *All of Us* Survey Artifacts. The final development will use the Delphi method to build consensus.

The initial development and construction of the index may contain some subjectivity in the mapping of the 'Deprivation Ontology' (each domain) to relevant survey artifacts, despite the use of the Jaccard Similarity Coefficient, because it was performed by an individual researcher. Further work is being done to rectify this via iterative consensus building with domain experts using the Delphi Method. Given the categorical nature of survey responses at the individual level, this study employed weighted multiple correspondence analysis, which is a well validated machine learning methodology that acts as a categorical corollary to primary component analysis. A final raw score was then calculated for each participant, and then the distribution of the scores broken into equal quintiles.

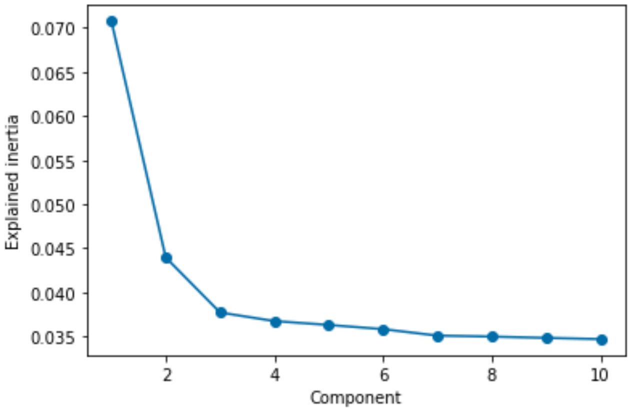


Figure 2: Inertia (variance) explained by each component. The weight of each relevant OMOP PPI concept used in combining socioeconomic domains was calculated using multiple correspondence analysis.

2.3 Validation

First, correlation between the ISDI and the Brokamp ADI was assessed between the mean of the ISDI aggregated at the 3-digit zip code level and the corresponding 3-digit Brokamp ADI score for each 3-digit zip code. Although coarse-grained, this is the most refined scale available within the workbench for geographic entity match. Second, for further validation, this study reconstructed an *All of Us* demonstration project that trained machine learning models adapted from a highly cited study that found racial bias in health algorithms[32]. This demonstration project is publicly available on the *All of Us* support center and researcher workbench and is designed to predict the health status of participants following the year of a participant's enrollment. The inputs to this model include:

- ➢ *Demographics: age at enrollment, gender, race and ethnicity, insurance, education*
- ➢ *Indicators for active chronic conditions at the enrollment year*

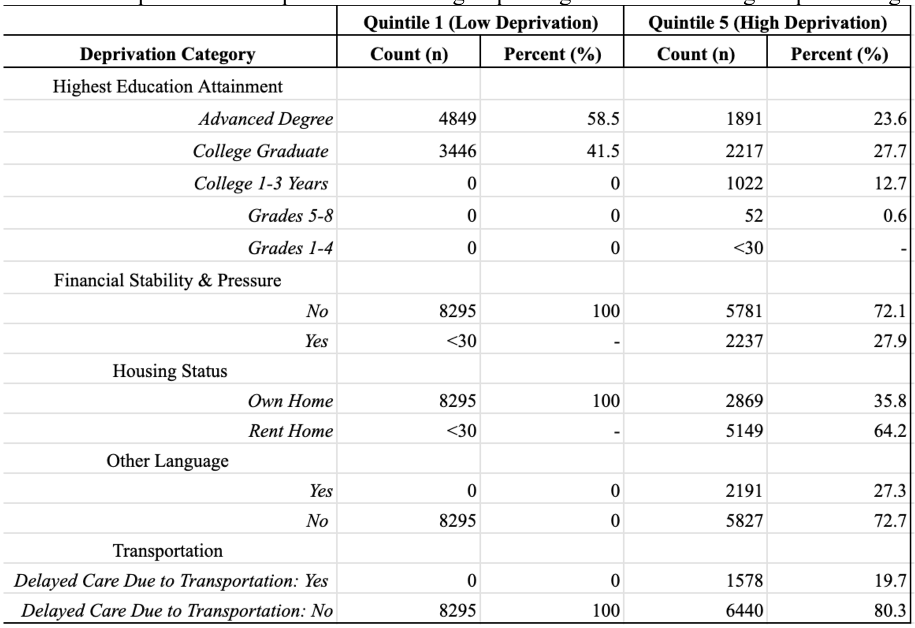
- *Biomarkers related to chronic diseases, including blood pressure, A1C, creatinine, hematocrit, and LDL as indicators: (normal, low, high)*
- *Medication: number of unique medications*

To better validate ISDI, we assess whether regularizing the data using stratified sampling based on ISDI quintile changes the model's performance and accuracy. The motivation for this specific application of ISDI is to assess the impact of de-biasing data network skewness as it relates to socioeconomic status. For example, each hospital that comprises of a data network encounters a different population structure, and to this end, AI models trained on these networks may benefit from regularization.

3 Results

Only participants who completed every relevant identified survey item were included, which was a total of 40,027 participants. The raw ISDI score assigned to each participant was normalized on a scale of 0-10. The distribution of raw ISDI scores was then split into quintiles, and each participant was assigned an ISDI quintile score from 1-5, with 5 being the most socioeconomically deprived group. Some of the breakdown of counts and percentages for the extreme deprivation limits, both high and low are shown in Table 1. These results follow expected trends: namely, there are greater percentages of high deprivation participants who indicated lower education attainment, less financial stability, less home ownership, lower income, language barriers, and delayed care due to transportation. The deviation was a small group of high earners in the high deprivation group.

Table 1. Select breakdown of counts and percentage of low and high deprivation per survey artifact response. This profile follows expected trends apart from a small group of high earners in the high deprivation group.



	Quintile 1 (Low Deprivation)		**Quintile 5 (High Deprivation)**	
Deprivation Category	**Count (n)**	**Percent (%)**	**Count (n)**	**Percent (%)**
Highest Education Attainment				
Advanced Degree	4849	58.5	1891	23.6
College Graduate	3446	41.5	2217	27.7
College 1-3 Years	0	0	1022	12.7
Grades 5-8	0	0	52	0.6
Grades 1-4	0	0	<30	-
Financial Stability & Pressure				
No	8295	100	5781	72.1
Yes	<30	-	2237	27.9
Housing Status				
Own Home	8295	100	2869	35.8
Rent Home	<30	-	5149	64.2
Other Language				
Yes	0	0	2191	27.3
No	8295	0	5827	72.7
Transportation				
Delayed Care Due to Transportation: Yes	0	0	1578	19.7
Delayed Care Due to Transportation: No	8295	100	6440	80.3

The correlation between the ISDI and the Brokamp ADI is assessed at a relatively large geographic area approximation. Due to privacy and reidentification concerns, *All of Us* currently only has the 3-digit zip code and its corresponding Brokamp ADI available for each participant. Again, this external data element is derived from ACS and an area approximation, and to this extent is more coarsely refined than ISDI which is constructed on an individual scale. This may explain the weak correlation seen in Figure 3. Finer area approximations may yield a stronger correlation, further validation may be explored to this extent.

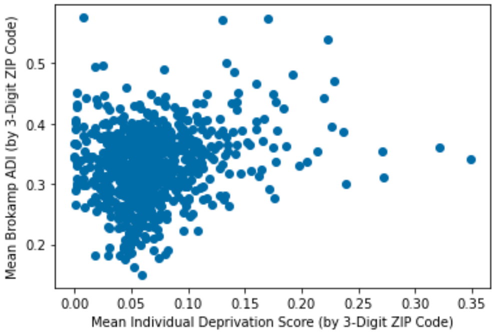


Figure 3. ISDI-ADI Correlation at the 3-digit zip code area entity. A comparison between the mean ISDI and the mean Brokamp ADI using a coarse geographic area approximation.

The last component of the analysis was the re-creation of an *All of Us* demonstration project derived from a highly cited study regarding racial bias in health algorithms[32]. The machine learning models were first trained normally using the standard train-test split. Then these same models were also trained with ISDI normalization (in this case: stratified sampling based on ISDI quintile). Table 2 demonstrates that although the accuracy of the models did not change, the AUC did decrease post-ISDI normalization for both logistic regression and random forest models. This may mean the model is becoming less discriminatory, which some literature indicates a more fair AI model across demographic groups[33]. Interestingly, when regularization was removed (L2), AUC increased regardless of ISDI normalization.

Table 2. Comparison of Accuracy and Area Under the Curve (AUC) Pre- and Post-ISDI Normalization. Notably While accuracy is maintained, AUC decreases post normalization, which is aligned with the notion that addressing biases may lead to fairer models which may also be less discriminatory across demographic groups.

	Pre ISDI Normalization		**Post ISDI Normalization**	
L2 Regularization	*LR Accuracy*	0.983	*LR Accuracy*	0.983
	LR AUC	0.581	*LR AUC*	0.562
No Regularization	*LR Accuracy*	0.983	*LR Accuracy*	0.983
	LR AUC	0.610	*LR AUC*	0.559
L2 Regularization with Varied Penalty Strength	*LR Accuracy*	0.983	*LR Accuracy*	0.983
	LR AUC	0.583	*LR AUC*	0.563
Random Forest	*RF Accuracy*	0.975	*RF Accuracy*	0.975
	RF AUC	0.887	*RF AUC*	0.841

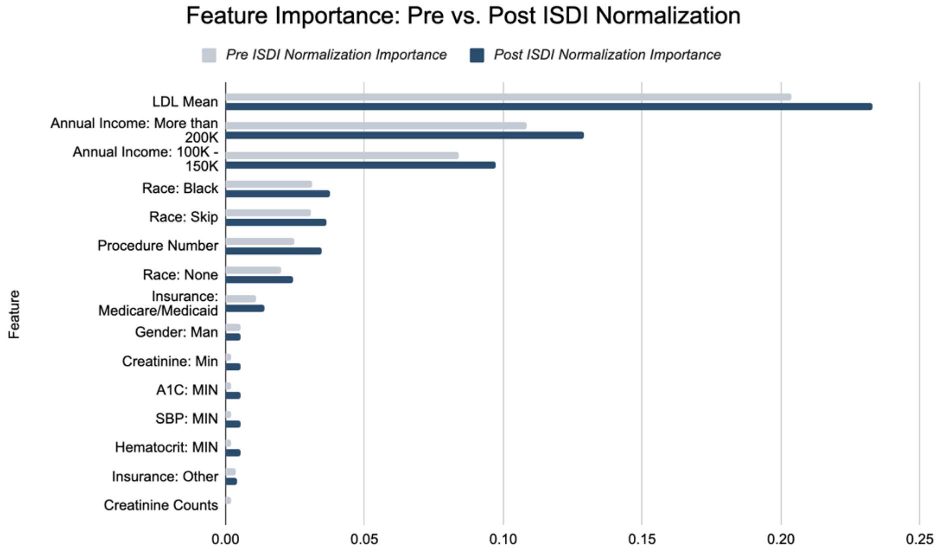


Figure 4. Comparison of the Pre- and Post- ISDI Normalization feature importance. Notably, although overall AUC was reduced, the importance of the certain features (e.g. 'Race: Black') went up.

Finally, Figure 4 compares the pre and post ISDI normalization feature importance, which can be interpreted in the context of the reduction of overall AUC from ISDI normalization. Features such as 'Race: Black' and 'Insurance: Medicare/Medicaid' became more important post normalization. The highlighted section indicates reordering of feature importance ranking pre and post normalization.

4 Discussion

This study builds on a body of work around indexing complex SDOH factors into composite area based deprivation measures. It extends this work into the development of an individual socioeconomic deprivation index (ISDI) constructed on a heterogenous data network designed to recruit UBR populations and capture its data diversity. Such an approach to indexing may have numerous unanticipated use cases in the context of precision medicine. The use of the ISDI to normalize a dataset for AI model training had tangible differences in AUC, while preserving accuracy, and simultaneously increasing the importance of features like 'Race: Black'. This can have multiple interpretations, and further validation will be needed across various models and contexts, nonetheless some literature has suggested such a trend can be interpreted in the context of making AI models more ethical[33].

A shortcoming of the initial development and construction of the index was the subjectivity in mapping the 'Deprivation Ontology' (each domain) to relevant survey artifacts, despite the use of the Jaccard Similarity Coefficient. Further work is being done to rectify this via iterative consensus building with domain experts using the Delphi Method.

Finally, this approach may be valuable in the light of growing health data networks and data linkages. For instance, distributed machine learning approaches such as federated learning models may hold significant potential for precision medicine at scale, but need to address the heterogeneity of the SDOH profile of various health data sources and account for non-independent and identically distributed (non-IID) datasets[34,35].

References

1. Marmot M, Friel S, Bell R, Houweling TAJ, Taylor S, on behalf of the Commission on Social Determinants of Health. Closing the gap in a generation: health equity through action on the social determinants of health. The Lancet. 2008;372(9650):1661-1669. doi:10.1016/S0140-6736(08)61690-6
2. Irwin A, Valentine N, Brown C, Loewenson R, Solar O, Brown H, Koller T, Vega J. The Commission on Social Determinants of Health: Tackling the Social Roots of Health Inequities. PLoS Med. 2006;3(6):e106. doi:10.1371/journal.pmed.0030106

3. Solar O, Irwin A. A conceptual framework for action on the social determinants of health. In: Figure A. Final form of the CSDH conceptual framework. Social Determinants of Health Discussion Paper 2 (Policy and Practice). World Health Organization; 2010. p.6.
4. Braveman P, Egerter S, Williams DR. The Social Determinants of Health: Coming of Age. Annu Rev Public Health. 2011;32:381-398. doi:10.1146/annurev-publhealth-031210-101218
5. Braveman PA, Egerter SA, Woolf SH, Marks JS. When do we know enough to recommend action on the social determinants of health? Am J Prev Med. 2011;40(1 Suppl 1):S58-66. doi: 10.1016/j.amepre.2010.09.026. PMID: 21146780
6. Braveman PA, Cubbin C, Egerter S, Williams DR, Pamuk E. Socioeconomic Disparities in Health in the United States: What the Patterns Tell Us. Am J Public Health. 2010;100(S1):S186-S196. doi:10.2105/AJPH.2009.166082. PMID: 20147693
7. Adler NE, Stewart J. Health disparities across the lifespan: meaning, methods, and mechanisms. Ann N Y Acad Sci. 2010;1186:5-23. doi:10.1111/j.1749-6632.2009.05337.x
8. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. Public Health Rep. 2014;129(Suppl 2):19-31. doi:10.1177/00333549141291S206
9. Phelan JC, Link BG, Tehranifar P. Social Conditions as Fundamental Causes of Health Inequalities: Theory, Evidence, and Policy Implications. J Health Soc Behav. 2010;51(1_suppl):S28-S40. doi:10.1177/0022146510383498
10. Krieger N. A glossary for social epidemiology. J Epidemiol Community Health. 2001;55(10):693-700. doi:10.1136/jech.55.10.693
11. Singh GK. Area deprivation and widening inequalities in US mortality, 1969-1998. Am J Public Health. 2003 Jul;93(7):1137-1143. doi:10.2105/ajph.93.7.1137. PMID: 12835199; PMCID: PMC1447923.
12. Kind AJ, Jencks S, Brock J, Yu M, Bartels C, Ehlenbach W, Greenberg C, Smith M. Neighborhood socioeconomic disadvantage and 30-day rehospitalization: a retrospective cohort study. Ann Intern Med. 2014 Dec 2;161(11):765-774. doi:10.7326/M13-2946. PMID: 25437404; PMCID: PMC4251560.
13. Kind AJH, Buckingham WR. Making Neighborhood-Disadvantage Metrics Accessible - The Neighborhood Atlas. N Engl J Med. 2018 Jun 28;378(26):2456-2458. doi:10.1056/NEJMp1802313. PMID: 29949490; PMCID: PMC6051533.
14. Flanagan BE, Gregory EW, Hallisey EJ, Heitgerd JL, Lewis B. A social vulnerability index for disaster management. J Homel Secur Emerg Manag. 2011;8(1). (This format is suitable for this particular journal.)
15. Brokamp C, Beck AF, Goyal NK, Ryan P, Greenberg JM, Hall ES. Material community deprivation and hospital utilization during the first year of life: an urban population-based cohort study. Ann Epidemiol. 2019;30:37-43. doi:10.1016/j.annepidem.2018.12.005.
16. Rosenzweig MQ, Althouse AD, Sabik L, Arnold R, Chu E, Smith TJ, ... Schenker Y. The association between area deprivation index and patient-reported outcomes in patients with advanced cancer. Health Equity. 2021;5(1):8-16.
17. Flanagan BE, Hallisey EJ, Adams E, Lavery A. Measuring community vulnerability to natural and anthropogenic hazards: The Centers for Disease Control and Prevention's Social Vulnerability Index. J Environ Health. 2018;80(10):34-36. PMID: 32327766; PMCID: PMC7179070.
18. Bakkensen LA, Fox-Lent C, Read LK, Linkov I. Validating resilience and vulnerability indices in the context of natural disasters. Risk Anal. 2017 May;37(5):982-1004. doi:10.1111/risa.12677. Epub 2016 Aug 30. PMID: 27577104.
19. KC M, Oral E, Straif-Bourgeois S, Rung AL, Peters ES. The effect of area deprivation on COVID-19 risk in Louisiana. PLoS One. 2020;15(12):e0243028.
20. Kurani S, McCoy RG, Inselman J, Jeffery MM, Chawla S, Rutten LJF, ... Shah ND. Place, poverty and prescriptions: a cross-sectional study using Area Deprivation Index to assess opioid use and drug-poisoning mortality in the USA from 2012 to 2017. BMJ Open. 2020;10(5):e035376.
21. U.S. EPA. Environmental Quality Index - Technical Report (2006-2010) (Final, 2020). U.S. Environmental Protection Agency, Washington, DC, EPA/600/R-20/367, 2020.
22. Maizlish N, Delaney T, Dowling H, Chapman DA, Sabo R, Woolf S, et al. California healthy places index: frames matter. Public Health Rep. 2019;134(4):354-362.
23. Richardson EA, Mitchell R, Shortt NK, Pearce J, Dawson TP. Developing summary measures of health-related multiple physical environmental deprivation for epidemiological research. Environ Plann A. 2010;42(7):1650-1668.

24. Arcaya MC, Tucker-Seeley RD, Kim R, Schnake-Mahl A, So M, Subramanian SV. Research on neighborhood effects on health in the United States: a systematic review of study characteristics. Soc Sci Med. 2016;168:16-29.
25. Hatef E, Kitchen C, Pandya C, Kharrazi H. Assessing Patient and Community-Level Social Factors; The Synergistic Effect of Social Needs and Social Determinants of Health on Healthcare Utilization at a Multilevel Academic Healthcare System. Journal of Medical Systems. 2023 Sep 1;47(1):95.
26. Tsou MH, Xu J, Lin CD, Daniels M, Embury J, Park J, Ko E, Gibbons J. Analyzing spatial-temporal impacts of neighborhood socioeconomic status variables on COVID-19 outbreaks as potential social determinants of health. Annals of the American Association of Geographers. 2023 Apr 21;113(4):891-912.
27. El-Sayed AM, Galea S. Temporal changes in socioeconomic influences on health: maternal education and preterm birth. American Journal of Public Health. 2012 Sep;102(9):1715-21.
28. Lopez J, Duarte G, Colombo RA, Ibrahim NE. Temporal Changes in Racial and Ethnic Disparities in the Utilization of Left Atrial Appendage Occlusion in the United States. The American Journal of Cardiology. 2023 Oct 1;204:53-63.
29. Cronin RM, Jerome RN, Mapes B, Andrade R, Johnston R, Ayala J, Schlundt D, Bonnet K, Kripalani S, Goggins K, Wallston KA, Couper MP, Elliott MR, Harris P, Begale M, Munoz F, Lopez-Class M, Cella D, Condon D, AuYoung M, Mazor KM, Mikita S, Manganiello M, Borselli N, Fowler S, Rutter JL, Denny JC, Karlson EW, Ahmedani BK, O'Donnell CJ; Vanderbilt University Medical Center Pilot Team, and the Participant Provided Information Committee. Development of the Initial Surveys for the All of Us Research Program. Epidemiology. 2019 Jul;30(4):597-608. doi: 10.1097/EDE.0000000000001028. PMID: 31045611; PMCID: PMC6548672.
30. Trinidad S, Brokamp C, Mor Huertas A, Beck AF, Riley CL, Rasnik E, Falcone R, Kotagal M. Use Of Area-Based Socioeconomic Deprivation Indices: A Scoping Review And Qualitative Analysis. Health Aff (Millwood). 2022 Dec;41(12):1804-1811. doi: 10.1377/hlthaff.2022.00482. PMID: 36469826.
31. Landscape of Area-Level Deprivation Measures and Other Approaches to Account for Social Risk and Social Determinants of Health in Health Care Payments Prepared for the Office of the Assistant Secretary for Planning and Evaluation (ASPE) at the U.S. Department of Health & Human Services by RAND Health Care, September 2022.
32. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019 Oct 25;366(6464):447-453. doi: 10.1126/science.aax2342. PMID: 31649194.
33. Pfohl S, Marafino B, Coulet A, Rodriguez F, Palaniappan L, Shah NH. Creating fair models of atherosclerotic cardiovascular disease risk. InProceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society 2019 Jan 27 (pp. 271-278).