Title:

Characterization and Racial Stratification of Social Determinants of Health for Individuals with Type 2 Diabetes as Recorded in Electronic Health Records: Implications for Artificial Intelligence Development

Corresponding author:
Polina Kukhareva, PhD, MPH
Assistant Professor, Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
421 Wakara Way, Suite 108
Salt Lake City, UT 84108
Email: polina.kukhareva@utah.edu
ORCID ID: 0000-0002-5576-1486

Authors

Polina V. Kukhareva, PhD, MPH,[1] Matthew J. O'Brien, MD, MSc,[2] Daniel C Malone, PhD,[3] Kensaku Kawamoto, MD, PhD, MHS,[1] Ramkiran Gouripeddi, MBBS, MS,[1] Deepika Reddy, MD,[4] Mingyuan Zhang, MS,[5] Vikrant G. Deshmuch, PhD,[1] JD, Julio C. Facelli, PhD[1]

[1] - Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, United States, [2] - Department of General Internal Medicine, Northwestern University, Chicago, IL, United States, [3] - Department of Pharmacotherapy, University of Utah, Salt Lake City, UT, United States, [4] - Diabetes and Endocrinology Center, University of Utah, Salt Lake City, Utah, United States, 5 - Department of Population Health Sciences, University of Utah, Salt Lake City, UT, 84108, United States

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. JCF and RG were partially funded by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UM1TR004409. MJO was partially funded by the National Institute of Diabetes and Digestive and Kidney Diseases under Award Number P30DK092949

Contributorship Statement

PVK, MJO, KK and JCF drafted the manuscript. PVK takes responsibility for the manuscript's content, including the data and analysis. Each author contributed substantially to the drafting or substantial revision of the paper. All authors contributed significantly to the study design, data interpretation, and manuscript writing. All authors also approved the paper for submission and agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated, resolved, and documented in the literature. PVK had full access to data. PVK conducted the statistical analyses.

Abstract

Background: Accurate documentation of social determinants of health (SDoH) in electronic health records (EHRs) is critical for developing equitable AI models for diabetes management. This study investigates SDoH data in a cross-institutional EHR database.

Methods: We analyzed neighborhood-level (i.e., social vulnerability index [SVI], Rural-Urban Community Area [RUCA]) and individual-level SDoH (e.g., preferred language, marital status, tobacco, alcohol, and substance use) within the Epic Cosmos database, focusing on adults diagnosed with T2D (E11.*) who had encounters between 2021 and 2023. We measured data completeness (i.e., the proportion of individuals who have a non-missing value) and the prevalence of non-canonical values (e.g., preference for language other than English) for each available SDoH variable.

Findings: The study included 12,696,680 individuals with T2D. SVI, RUCA and preferred language were available for all individuals, while marital status, and smoking data were available for over 90%. However, financial needs, interpersonal violence, social activity, and physical activity were present in EHRs for 7.6%-24.6% of the population depending on race/ethnicity. Minority groups experienced lower data completeness and higher burden of non-canonical values compared to White individuals.

Interpretation: Neighborhood-level and some individual-level SDoH have potential for use in AI development and evaluation. Other SDoH data cannot be used without additional analysis to address high amounts of missing data. Significant disparities in completeness exist across racial/ethnic groups. Addressing these data gaps may require government and payer mandates, standardized SDoH screening tools, and personnel training.

Highlights

1. This study examined social determinants of health (SDoH) data for adults with type 2 diabetes in a cross-institutional electronic health record (EHR) database to support equitable AI model development.
2. Neighborhood-level SDoH data and some individual-level SDoH data (individual-level SDoH (i.e., race/ethnicity, preferred language, marital status) were highly complete.
3. Disparities in SDoH data completeness by race/ethnicity underscore the need for standardized SDoH documentation.

Keywords

social determinants of health; electronic health records, type 2 diabetes

Background

Type 2 diabetes (T2D) is a chronic metabolic condition that affected 38.4 million U.S. adults in 2021 and is expected to affect 60.6 million by 2060.[1,2] To address this public health crisis, the use of artificial intelligence (AI) for managing T2D in clinical settings is on the rise.[3] For example, several models were recently developed to provide personalized T2D medication recommendations.[4] Electronic health records (EHRs), with almost universal adoption and growing data interoperability, serve as an essential source of data for clinical AI algorithms.[4]

While use of AI models based on EHR data is an promising and exciting development,[4] there is a growing concern that AI models may exacerbate existing healthcare disparities by underperforming in populations that are already vulnerable.[5–10] Biases inherent in the training data, such as lack of representation and informativeness for certain demographic groups, can lead to proposing suboptimal treatment plans.[11] Despite these concerns, there has been relatively little effort to systematically identify and mitigate bias in clinical ML models,[12] making it crucial to address these issues to ensure equitable healthcare for all populations.

The prevalence of T2D is nearly two times higher in racial and ethnic non-White groups than among White Americans, highlighting health inequities that may be shaped in part by the social determinants of health (SDoH), the conditions in which people are born, grow, live, work, and age.[13–16] These disparities make it especially important to ensure that the emerging AI-algorithms for clinical management of T2D are fair. AI fairness refers to ensuring that AI models perform equitably across different demographic groups. High completeness of SDoH variables is essential to evaluate fairness of emerging AI algorithms.

EHRs could potentially become a primary source of SDoH data for validation of AI algorithms.[17] Several healthcare stakeholders, including CMS and the Joint Commission, mandate the collection of certain SDoH data, and many health systems are voluntarily striving to collect this information. Epic, which is used by approximately half of all ambulatory medical providers, is prioritizing the collection of SDoH data,[18] and there are many ongoing efforts to standardize the collection of SDoH data in EHRs using validated questionnaires.[19–21] Additionally, data interoperability standards are starting to be used to store and exchange SDoH data in a consistent way.[22] However, while early reports indicate that the utility of such SDoH data is limited by low completeness,[20] to the authors knowledge, no comprehensive studies stratified by race/ethnicity in multi-institutional data repositories have been reported for individuals with T2D.

This study aimed to address whether SDoH variables are sufficiently present in EHRs to enable accurate AI fairness evaluation in a very large multi-institutional EHR data repository, Epic Cosmos.[23]

Methods

Setting

Data used in this study came from Epic Cosmos,[24] a community collaboration of health systems representing over 251 million individual records from over 1,400 hospitals and 32,500 clinics. Epic Cosmos includes aggregated data from a substantial percentage of the U.S. population. Data extraction was completed on June 14, 2024.

Study Population

We included adults (age 18 years or older at the data extraction date) who had an office visit, telemedicine encounter, surgery visit, or lab visit in 2021-2023 and had a T2D diagnosis code (E11.*) as an encounter diagnosis, billing final diagnosis or active in the problem list in 2015-2023. Combined race/ethnicity was aggregated into American Indian or Alaskan Native (AI/AN), Asian, Black or African American (Black), Hispanic or Latino (Hispanic), Native Hawaiian or Other Pacific Islander (NH/PI), and White. Non-white individuals were defined as those who reported race/ethnicity other than White or Caucasian. Persons identifying with more than one racial category were excluded from the analysis of individual-level SDoH.

Neighborhood-level SDoH

Unlike individual-level SDoH variables, which are specific to the individual, neighborhood-level SDoH are based on the individual's residence and may be less precise for an individual. We analyzed two neighborhood-level SDoH variables: Social Vulnerability Index (SVI) and Rural-Urban Community Area (RUCA) codes. The SVI helps public health officials and planners prepare communities for emergencies like severe weather, disease outbreaks, or chemical exposure. It is based on 16 U.S. census variables and includes four main categories: socioeconomic status, racial and ethnic minority status, household characteristics, and housing and transportation.[25] SVI scores range from 0 (less vulnerable) to 1 (more vulnerable). RUCA codes categorize U.S. census tracts by urbanization, population density, and commuting patterns. In the EHR databases, neighborhood-level variables are imputed from the individual address, which is widely available.

Individual-level SDoH

Individual-level SDoH can be documented in the EHR by clinical providers and medical support staff. Collection of SDoH data is standardized using integrated SDoH displays (**Figure 1**). We included all available SDoH variables that are currently available in Epic Cosmos. Available variables included preferred language, marital status, 12 variables related to the use of substances with potential for misuse and addiction (tobacco, alcohol, psychoactive drugs), 6 variables related to resource needs (financial resource strain, food scarcity, food worry, medical transportation needs, non-medical transportation needs, and housing instability), 4 variables related to interpersonal violence (emotional, fear, physical abuse, sexual abuse), 2 variables related to physical activity (days per week, minutes per session), 5 variables related to social connections (church, get together, meetings, membership, phone) and 1 attribute related to stress. For individuals who had multiple values recorded, we used the last value collected prior to the data extraction date.

Study Measures

This study evaluated completeness and prevalence of non-canonical values for each SDoH variable. Completeness of an SDoH variable was defined as the proportion of individuals who have a non-missing value for this attribute. We also evaluated the completeness of detailed information for individuals who reported use of substances such as smoking history, alcohol use, and substance use. For example, we evaluated the completeness of detailed smoking history (pack-years and years smoked) for individuals who reported smoking. Prevalence of non-

canonical answers referred to answers different from the canonical value. We defined canonical answers as those expected by the majority class (White), for instance English language:

- Preferred language: English
- Marital status: Married
- Tobacco Use, Smokeless; Tobacco Use, Smoking, Alcohol Use, and Substance Use: Never used
- Food Scarcity; Food Worry: Never true
- Transportation needs, housing instability, interpersonal violence: No issues
- Financial Resource Strain: Not hard at all
- Physical activity: More than 0 days and 0 minutes of activity
- Social activity: Any level of activity other than never
- Stress: Not at all stressed
- Cigarette smoking: Less than 20 pack-years and less than 1 pack per day
- Alcohol use: Less than 1 drink per day, less than 2 drinks per week, fewer than 3 standard drinks, and less than 1 binge per month
- Drug misuse: Once or less
- Substance abuse: Use of drugs other than the following five drugs causing most overdose deaths (fentanyl, methamphetamine, cocaine, heroin, and oxycodone).[26]

Statistical Analysis

Descriptive characteristics were summarized using N (%) for sex and RUCA. Mean (standard deviation) were used for age and SVI. Results with less than 10 observations are reported as <10. Completeness of SDoH data and prevalence of non-canonical values were estimated using logistic regression adjusting for individual age and sex at the data extraction date. The reference racial category was White.

Results

Table 1 displays the demographic characterization of the 12,696,680 individuals with T2D. This is about a third of all individuals within the U.S. with T2D. The average age was 65.12 (SD: 14.77) years, and 50.1% were female.

Neighborhood-level SDoH (SVI and RUCA) are also summarized in Table 1. SVI was available for 100% of individuals and RUCA was available for 99.4% of individuals.

Completeness of SDoH variables by race/ethnicity is summarized in **Figure 2** and **Supplement Table S1**. **Figure 3** and **Supplement Table S2** summarize data related to SDoH by race/ethnicity group. Below we discuss the findings for each of the 32 variables evaluated.

Preferred Language: This variable was universally available, without any race/ethnicity differentiation. The Hispanic group reported a language preference different than the white majority class (English). Asians and NHPIs had a substantial non-White that preferred languages other than English.

Marital Status: Marital status was reported for 93.6-97.7% of the cohort depending on race/ethnicity. The non-canonical values were reported for between 29.0-62.2% of the racial subgroups. Black (OR: 1.41 [95% CI: 1.39-1.44]) and AI/AN (OR: 1.59 [95% CI: 1.56-1.62])

individuals were more likely not to be married compared to White individuals, while Asian individuals were less likely not to be married (OR: 0.52 [95% CI: 0.52-0.53]).

Tobacco Use: Data on smoking and smokeless tobacco use were reported for 87.8-92.3% of the subgroups and were available slightly more often for individuals classified as White. Forty-eight percent of T2D individuals reported a positive history of smoking and 6% reported using smokeless tobacco. Among smokers, detailed smoking information was available for over 60% of White group, but less for individuals from racial non-White groups. Fifty five percent of AI/AN with available smoking information reported that they currently smoke of smoked in the past. Asian (OR: 0.52 [95% CI: 0.51-0.53]) and Hispanic (OR: 0.60 [95% CI: 0.60-0.61]) individuals had significantly lower odds of smoking compared to White individuals. Sixty seven percent of White persons had cumulative smoking history of more than 20 pack-years. Among those with non-White race/ethnicity, fewer individuals reported smoking more than 20 pack-years.

Alcohol Use: Among persons identified as White, completeness on alcohol use was 80.0% [95% CI: 80.0-80.1]. Availability of alcohol exposure was lower for AI/AN (OR: 0.95 [95% CI: 0.93-0.97]), Asian (OR: 0.71 [95% CI: 0.71-0.72]), and Hispanic (OR: 0.79 [95% CI: 0.79-0.8]) individuals and higher for Black (OR: 1.24 [95% CI: 1.23-1.25]) and NH/PI (OR: 1.1 [95% CI: 1.07-1.13]) individuals compared to White individuals (**Table S1**). AI/AN (OR: 0.54 [95% CI: 0.52-0.55]), Asian (OR: 0.49 [95% CI: 0.49-0.50]), Black (OR: 0.77 [95% CI: 0.77-0.78]), Hispanic (OR: 0.56 [95% CI: 0.55-0.56]), and NH/PI (OR: 0.55 [95% CI: 0.53-0.57]) individuals were less likely to consume alcohol compared to White individuals (**Table S2**). Among those who drink, AI/AN reported higher alcohol consumption and more binge drinking.

Substance Use: Among persons identified as White, the data completeness was 58.4 [95% CI: 58.4-58.5]. Availability of reported substance use was lower for AI/AN, Asian, and Hispanic individuals and higher for Black and NH/PI individuals compared to White individuals. Further, if substance use was reported, Black individuals were more likely to be asked about the type of substance used (OR = 1.35 [95% CI: 1.3-1.39]). Among those with available data, AI/AN (OR: 1.64 [95% CI: 1.58-1.71]) and Black individuals (OR: 1.24 [95% CI: 1.23-1.25]) had higher odds of substance use compared to White individuals, while Asian (OR: 0.19 [95% CI: 0.18-0.2]) and Hispanic (OR: 0.57 [95% CI: 0.56-0.58]) individuals had lower odds. Minority individuals had higher odds of using the five drugs that lead to most overdose deaths.

Patient needs: Data completeness for food scarcity, food worry, transportation needs, housing instability and financial resource strain was recorded for 7.6% to 24.6% of individuals across race/ethnicity groups. Among those with available data, AI/AN, Black NHPI and Hispanic individuals had higher unmet needs compared to persons identified as White. Asian individuals had lower level of unmet financial needs. AI/AN, Black, Hispanic, and NH/PI individuals experienced higher odds of housing instability and faced greater financial resource strain compared to persons identified as White.

Interpersonal Violence: Data completeness for interpersonal violence ranged from 6.2% to 11.3%. From those with complete data, less than 1% reported experiencing interpersonal violence. AI/AN individuals reported slightly higher levels of physical abuse. Asian individuals reported slightly lower rates. AI/AN individuals had similar odds of experiencing emotional violence, fear-based violence, and physical abuse compared to White individuals, while Asian, Black, Hispanic, and NH/PI individuals had lower odds of this exposure.

Physical activity: Data completeness for physical activity was around from 8.2% to 14.4%. Completeness of physical activity data was lower for AI/AN, NH/PI, Asian, and Hispanic individuals. Asian individuals reported high rates of physical activity compared to other groups. More AI/AN and Hispanic individuals reported not exercising. Hispanic and AI/AN individuals were more likely to report 0 minutes of physical activity per session compared to White individuals. In contrast, Asian individuals were less likely to report zero minutes of physical activity.

Social Activity: Complete data for social activity was available for 5.8% to 11.2% of individuals. All individuals from non-White groups had less complete data than persons identified as White. Black, Asian, Hispanic, and NH/PI individuals were less likely to never attend church compared to White individuals, indicating higher church participation among these groups. Hispanic individuals were more likely to have no meetings, and no memberships documented.

Stress: Data completeness for stress ranged from 7.4% for NH/PI to 13.5% among White individuals. All individuals of non-White race/ethnicity had less complete data than persons identified as White. Among those with available data, non-White individuals reported less stress than persons identified as White.

Discussion

This study examines the completeness of Social Determinants of Health (SDoH) data for individuals with T2D in the Epic Cosmos database. Neighborhood-level SDoH variables, such as the SVI and RUCA, are universally available due to the necessity of addresses for administrative purposes. Additionally, individual-level SDoH data—including race/ethnicity, sex, age, preferred language, marital status, and substance use (tobacco, alcohol, and drugs)—are broadly documented. The widespread availability of substance use history is particularly promising, given its significant influence on T2D progression.[27–30] These data are crucial for identifying individuals eligible for targeted interventions, thereby helping to reduce complications and mortality through treatment, cessation, and screening programs. However, the completeness of substance use data varies by substance type and race/ethnicity, ranging from 43.9% to 92.3%, suggesting that data augmentation methods may be necessary.[31,32] These findings suggest that equity analyses using SDoH variables are feasible, though caution is warranted.

Conversely, several variables are not yet suitable for AI fairness research due to limited completeness. Data on resource needs, interpersonal violence, physical activity, social connections, and stress are available for only 5.6%-24.6% of individuals, depending on the attribute and race/ethnicity. The low completeness rates for these variables are concerning, as they could provide critical insights into the lived experiences of individuals with T2D, significantly impacting disease management and outcomes. The lack of complete data complicates the understanding of patient challenges and may result in less effective interventions. Significant barriers remain to systematically collecting these data in EHRs. Healthcare teams are already burdened with data collection and entry, and the time required to ask SDoH-related questions and address non-canonical results is substantial. Moreover, reluctance among healthcare teams to inquire about SDoH issues is often due to a lack of resources to address identified problems. Additionally, the frequency of data collection and its impact on outcomes remain unclear.

We observed systemic under-collection of SDoH data in Non-White racial/ethnic groups, which coincided with a higher burden of non-canonical SDoH values compared to White individuals. AI models trained on biased data may inadvertently reinforce existing biases, compromising performance in Non-White groups.[33] Insufficient and biased SDoH data may also lead to AI systems that fail to accurately represent these groups, perpetuating health disparities and leading to inequities in resource allocation.[33] Addressing this under-collection is critical for AI fairness, requiring better data collection, augmentation, and continuous bias monitoring. Further research should focus on mitigating varying levels of SDoH data completeness across racial/ethnic groups using statistical and AI methods for data augmentation.[31,32]

SDoH data collection might also be influenced by systemic racism. For instance, Black and Native Hawaiian/Pacific Islander (NH/PI) individuals were more likely to have alcohol and substance use recorded, potentially reflecting differences in data collection practices or clinician biases. Despite higher data completeness for alcohol consumption among Black individuals, reported consumption rates were lower compared to White individuals. This disparity may stem from stereotypes that view Black individuals through a lens of violent behavior and substance abuse. The availability of alcohol and substance use data, coupled with the lack of comprehensive data on other SDoH, may lead to an incomplete understanding of Black and NH/PI individuals' health needs and less effective interventions.

Our study also identified significant disparities in the prevalence of non-canonical values across different racial and ethnic groups, underscoring the need for targeted interventions to address these inequities. For example, American Indian/Alaska Native (AI/AN) individuals, who experience higher transportation needs, may benefit from programs specifically designed to address and mitigate transportation barriers in these communities. These findings can inform policy decisions and resource allocation, ensuring that interventions and healthcare services are directed toward the populations most in need, ultimately contributing to the development of more effective and sustainable health programs.

Limitations

This study has several limitations. First, it relies on retrospective EHR data, which are prone to documentation biases and missing information. While the Epic Cosmos database offers a comprehensive source of data from various healthcare environments, the generalizability of our findings may be constrained to individuals treated at institutions using the Epic EHR system. However, given that Epic Cosmos includes data from more than half of the U.S. population, we believe our results are broadly applicable. Second, the definition of SDoH is still evolving, and some of the variables studied (e.g., substance use) might not be included as SDoH by some stakeholders. Third, our study was limited to SDoH variables available in Epic Cosmos. Some essential SDoH variables are not currently collected in EHRs or are unavailable in the Epic Cosmos database. Further, we looked at only SDoH variables recorded in structured, while it is possible to extract SDoH from clinical notes using natural language processing and text mining methods,[34] notes are not available at Epic Cosmos currently. Fourth, there is a high likelihood that SDoH data are collected differently across healthcare organizations, potentially affecting the validity of the findings. Lastly, we did not use the SUrveillance, PREvention, and ManagEment of Diabetes Mellitus (SUPREME-DM) algorithm, which could have identified more individuals with T2D. However, we focused on individuals with documented T2D diagnoses, using more restrictive inclusion criteria.

Conclusion

This study highlights the feasibility and importance of systematically collecting SDoH data for individuals with T2D. The widespread availability of neighborhood-level SDoH variables and some individual-level SDoH variables, such as race, ethnicity, legal sex, preferred language, marital status, and substance use, demonstrates the potential for EHR data to support AI equity research. However, the study also identifies concerning gaps in the completeness of critical individual-level SDoH variables such as resource needs, interpersonal violence, physical activity, social connections, and stress. The findings underscore the need to recognize the value of SDoH in healthcare data systems and the importance of equitable data collection.

Data Availability Statement

Authors have no acknowledgments.

Disclosures

Authors have no disclosures.

Data Availability Statement

The data underlying this article cannot be shared publicly due to Epic Cosmos policies.

References

1. Centers for Disease Control and Prevention. *National Diabetes Statistics Report.*; 2023.
2. Lin J, Thompson TJ, Cheng YJ, et al. Projection of the future diabetes burden in the United States through 2060. *Popul Health Metr*. 2018;16(1). doi:10.1186/S12963-018-0166-4
3. Sheng B, Pushpanathan K, Guan Z, et al. Artificial intelligence for diabetes care: current and future prospects. *Lancet Diabetes Endocrinol*. 2024;12(8):569-595. doi:10.1016/S2213-8587(24)00154-2
4. Fujihara K, Sone H. Machine Learning Approach to Drug Treatment Strategy for Diabetes Care. *Diabetes Metab J*. 2023;47(3):325. doi:10.4093/DMJ.2022.0349
5. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342
6. Madhusoodanan J. Is a racially-biased algorithm delaying health care for one million Black people? *Nature*. 2020;588(7839):546-547. doi:10.1038/D41586-020-03419-6
7. Ledford H. Millions of black people affected by racial bias in health-care algorithms. *Nature*. 2019;574(7780):608-609. doi:10.1038/D41586-019-03228-6
8. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern Med*. 2018;178(11):1544. doi:10.1001/JAMAINTERNMED.2018.3763

9. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf*. 2019;28(3):231-237. doi:10.1136/BMJQS-2018-008370

10. Dorr DA, Adams L, Embí P. Harnessing the Promise of Artificial Intelligence Responsibly. *JAMA*. 2023;329(16):1347-1348. doi:10.1001/JAMA.2023.2771

11. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med*. 2018;169(12):866. doi:10.7326/M18-1990

12. Taber P, Armin JS, Orozco G, et al. Artificial Intelligence and Cancer Control: Toward Prioritizing Justice, Equity, Diversity, and Inclusion (JEDI) in Emerging Decision Support Technologies. *Curr Oncol Rep*. 2023;25(5):387-424. doi:10.1007/S11912-023-01376-7

13. NEJM Catalyst. Social Determinants of Health (SDOH). doi:doi:10.1056/CAT.17.0312

14. Hill-Briggs F, Adler NE, Berkowitz SA, et al. Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes Care*. 2020;44(1):258-279. doi:10.2337/dci20-0053

15. Hill-Briggs F, Ephraim PL, Vrany EA, et al. Social Determinants of Health, Race, and Diabetes Population Health Improvement: Black/African Americans as a Population Exemplar. *Curr Diab Rep*. 2022;22(3):117-128. doi:10.1007/s11892-022-01454-3

16. Hill-Briggs F, Fitzpatrick SL. Overview of Social Determinants of Health in the Development of Diabetes. *Diabetes Care*. 2023;46(9):1590-1598. doi:10.2337/dci23-0001

17. Chen M, Tan X, Padman R. Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *J Am Med Inform Assoc*. 2020;27(11):1764. doi:10.1093/JAMIA/OCAA143

18. Epic's EHR Optimization Mitigates SDOH, Promotes Care Coordination | TechTarget. Accessed August 1, 2024. https://www.techtarget.com/searchhealthit/news/366579171/Epics-EHR-Optimization-Mitigates-SDOH-Promotes-Care-Coordination

19. Craven CK, Highfield L, Basit M, et al. Toward standardization, harmonization, and integration of social determinants of health data: A Texas Clinical and Translational Science Award institutions collaboration. *J Clin Transl Sci*. 2024;8(1):e17. doi:10.1017/cts.2024.2

20. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. *J Am Med Inform Assoc*. 2021;29(1):187-196. doi:10.1093/JAMIA/OCAB199

21. Dullabh P, Hovey L, Leaphart D, Chiao A, Heaney-Huls K. *Expanding Social Determinants of Health Data across PCORnet® Clinical Research Networks*.; 2022.

22. Phuong J, Zampino E, Dobbins N, et al. Extracting Patient-level Social Determinants of Health into the OMOP Common Data Model. *AMIA Annual Symposium Proceedings*. 2021;2021:989. Accessed August 19, 2024. /pmc/articles/PMC8861735/

23. Tarabichi Y, Frees A, Honeywell S, et al. The Cosmos Collaborative: A Vendor-Facilitated Electronic Health Record Data Aggregation Platform. *ACI open*. 2021;5(1):e36. doi:10.1055/S-0041-1731004

24. Epic Cosmos. Accessed August 6, 2024. https://cosmos.epic.com/about/

25. CDC/ATSDR Social Vulnerability Index (SVI).

26. Spencer MR, Warner M, Cisewski JA, et al. Estimates of Drug Overdose Deaths Involving Fentanyl, Methamphetamine, Cocaine, Heroin, and Oxycodone: United States, 2021. Published online 2021. Accessed July 24, 2024. https://www.cdc.gov/nchs/products/index.htm.

27. Campagna D, Alamo A, Di Pino A, et al. Smoking and diabetes: dangerous liaisons and confusing relationships. *Diabetology & Metabolic Syndrome 2019 11:1*. 2019;11(1):1-12. doi:10.1186/S13098-019-0482-2

28. Mayl JJ, German CA, Bertoni AG, et al. Association of Alcohol Intake With Hypertension in Type 2 Diabetes Mellitus: The ACCORD Trial. *J Am Heart Assoc*. 2020;9(18). doi:10.1161/JAHA.120.017334

29. van de Wiel A. Diabetes mellitus and alcohol. *Diabetes Metab Res Rev*. 2004;20(4):263-267. doi:10.1002/DMRR.492

30. Ojo O, Wang XH, Ojo OO, Ibe J. The Effects of Substance Abuse on Blood Glucose Parameters in Patients with Diabetes: A Systematic Review and Meta-Analysis. *Int J Environ Res Public Health*. 2018;15(12):2691. doi:10.3390/IJERPH15122691

31. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol*. 2019;110:63-73. doi:10.1016/J.JCLINEPI.2019.02.016

32. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, et al. Missing data and multiple imputation in clinical epidemiological research. *Clin Epidemiol*. 2017;9:157-166. doi:10.2147/CLEP.S129785

33. Chin MH, Afsar-Manesh N, Bierman AS, et al. Guiding Principles to Address the Impact of Algorithm Bias on Racial and Ethnic Disparities in Health and Health Care. *JAMA Netw Open*. 2023;6(12):E2345050. doi:10.1001/JAMANETWORKOPEN.2023.45050

34. Brown JR, Ricket IM, Reeves RM, et al. Information Extraction From Electronic Health Records to Predict Readmission Following Acute Myocardial Infarction: Does Natural Language Processing Using Clinical Notes Improve Prediction of Readmission? *J Am Heart Assoc*. 2022;11(7). doi:10.1161/JAHA.121.024198

Figure legends

Figure 1. SDoH Data Captured and Displayed in the EHR.

Figure 2. Completeness of SDoH Variables Among Patients with T2D by Race/Ethnicity. A Percentage of complete records. B. Odds ratios compared to White individuals.

All values were calculated using logistic regression and adjusted for age and sex.
Abbreviations: AI/AN – American Indian or Alaskan Native, Black – Black or African American, Hispanic – Hispanic or Latino, NH/PI – Native Hawaiian or Other Pacific Islander

Figure 3. Presence of Issues Among Patients with T2D by Race/Ethnicity. A Percentage of records with issues. B. Odds ratios compared to White individuals.

All values were calculated using logistic regression and adjusted for age and sex. Odds ratios above 4 are removed.
Abbreviations: AI/AN – American Indian or Alaskan Native, Black – Black or African American, Hispanic – Hispanic or Latino, NH/PI – Native Hawaiian or Other Pacific Islander

Tables

Table 1. Patient Demographics and Neighborhood-level SDoH by Race/Ethnicity

Variable	White	AI/AN	Asian	Black	Hispanic	NH/PI	More than one race	Other	Unknown	Overall
Eligible Patients, N	7,350,527	54,007	408,800	2,195,298	1,447,925	32,965	807,237	159,099	240,822	12,696,680
Age, years	67.51 (14)	60.98 (14.81)	64.81 (14.79)	62 (15.03)	59.43 (15.35)	59.9 (14.67)	63.49 (15)	63.23 (15.03)	63.75 (14.82)	65.12 (14.77)
Legal Sex, N (%)										
- Female	3,456,336 (47.02)	28,566 (52.89)	205,753 (50.33)	1,263,230 (57.54)	766,032 (52.91)	17,391 (52.76)	439,875 (54.49)	74,269 (46.68)	108,666 (45.12)	6,360,118 (50.09)
- Male	3,893,590 (52.97)	25,437 (47.1)	202,984 (49.65)	931,875 (42.45)	681,686 (47.08)	15,568 (47.23)	367,249 (45.49)	84,776 (53.29)	131,658 (54.67)	6,334,823 (49.89)
- Other	18 (0)	<10	<10	<10	<10	<10	<10	<10	<10	46 (0)
- Unknown	583 (0.01)	<10	61 (0.01)	190 (0.01)	198 (0.01)	<10	105 (0.01)	51 (0.03)	495 (0.21)	1693 (0.01)
SVI										
- Socioeconomic Status	0.46 (0.28)	0.61 (0.29)	0.39 (0.29)	0.66 (0.28)	0.65 (0.28)	0.45 (0.28)	0.53 (0.3)	0.49 (0.3)	0.51 (0.3)	0.52 (0.29)
- Racial and Ethnic Minority Status	0.58 (0.24)	0.69 (0.2)	0.81 (0.16)	0.78 (0.16)	0.86 (0.15)	0.83 (0.16)	0.71 (0.2)	0.79 (0.18)	0.75 (0.21)	0.67 (0.24)
- Household Composition	0.49 (0.28)	0.6 (0.29)	0.31 (0.24)	0.58 (0.28)	0.5 (0.27)	0.4 (0.25)	0.51 (0.29)	0.4 (0.28)	0.45 (0.28)	0.5 (0.28)
- Housing Type and Transportation	0.56 (0.27)	0.69 (0.25)	0.59 (0.28)	0.69 (0.24)	0.71 (0.24)	0.67 (0.26)	0.61 (0.26)	0.61 (0.27)	0.62 (0.27)	0.6 (0.27)
RUCA, N (%)										
- Metropolitan	5,672,838 (77.18)	32,841 (60.81)	392,611 (96.04)	1,982,398 (90.3)	1,330,892 (91.92)	28,795 (87.35)	699,069 (86.6)	146,349 (91.99)	209,958 (87.18)	10,495,751 (82.67)
- Micropolitan	882188 (12)	8278 (15.33)	9714 (2.38)	119755 (5.46)	70632 (4.88)	2863 (8.68)	62211 (7.71)	5442 (3.42)	15989 (6.64)	1177072 (9.27)
- Small town	454427	5720	2677	61438	23879	668	28208	2128	8241	587386

	(6.18)	(10.59)	(0.65)	(2.8)	(1.65)	(2.03)	(3.49)	(1.34)	(3.42)	(4.63)
- Rural	297393 (4.05)	7053 (13.06)	918 (0.22)	21755 (0.99)	10334 (0.71)	443 (1.34)	14467 (1.79)	1163 (0.73)	5210 (2.16)	358736 (2.83)
- Unable to calculate	43681 (0.59)	115 (0.21)	2880 (0.7)	9952 (0.45)	12188 (0.84)	196 (0.59)	3282 (0.41)	4017 (2.52)	1424 (0.59)	77735 (0.61)

Values are expressed as mean (S.D.) unless otherwise specified.
Abbreviations: SVI – social vulnerability index, AI/AN – American Indian or Alaskan Native, Black – Black or African American, Hispanic – Hispanic or Latino, NH/PI – Native Hawaiian or Other Pacific Islander, RUCA - Rural-Urban Community Area.

History

GENERAL
Medical
Surgical
Family

SOCIAL DETERMINANTS
Substance & Sexual ...
E-cigarette/Vaping
Socioeconomic
Social Determinants
Social Documentation

SPECIALTY
Birth

Review Social Determinants

Social Determinants of Health

Expand All Collapse Al

Tobacco Use
Jan , 2024: Low Risk

Alcohol Use
Apr , 2023: Not At Risk

Depression
Apr , 2023: Not at risk

Personal Safety
Not on file

Food Insecurity
Aug , 2023: Low Risk
May , 2023: Food Insecurity Present

Housing Stability
May , 2023: Low Risk

Utilities
Not on file

Transportation Needs
Aug , 2023: Low Risk

Financial Resource Strain
May , 2023: High Risk

Medical Cost Burden
Apr 2024: High Risk

May , 2023: High Risk

Overall Financial Resource Strain (CARDIA)

Difficulty of Paying Living Expenses
Hard

Jan , 2024: High Risk

SINCERE Medical Cost Burden

Did not see a doctor because it costs too much in the past month
Yes

Did not take medications to save money in the past month
Yes

Employment Status
Not on file

Caregiver Impact
Not on file

Social Connections
Not on file

Stress
Not on file

Physical Activity
Not on file