Landscape of racial and ethnic health disparities in the All of Us Research Program

Vincent Lam
NIMHD
Shivam Sharma
Georgia Institute of Technology
I. King Jordan
Georgia Institute of Technology
Leonardo Mariño-Ramírez (marino@nih.gov)
NIMHD https://orcid.org/0000-0002-5716-8512

Research Article

Keywords: All of Us, health disparities

Posted Date: November 17th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3621210/v1

Abstract

The All of Us Research Program (*All of Us*) is an initiative led by the United States National Institutes of Health (NIH) whose goal is to advance research on personalized medicine and health equity through the collection of genetic, environmental, demographic, and health data from volunteer participants who reside in the United States (US). The program's emphasis on recruiting a diverse participant cohort makes *All of Us* an effective platform for investigating racial health disparities. However, to our knowledge, there have been no attempts to catalog the landscape of racial and ethnic health disparities that exist in the *All of Us* participant cohort. In this work, we analyzed participant electronic health record (EHR) data to identify the diseases and disease categories in the *All of Us* cohort for which racial and ethnic prevalence disparities can be observed. In conjunction with these analyses, we developed the US Health Disparities Browser as an interactive web application that enables users to visualize differences in race- and ethnic-group specific prevalence estimates for 1,755 different diseases: https://usdisparities.biosci.gatech.edu/. The web application features a catalog of all diseases represented in the browser, which can be sorted by overall prevalence as well as the variance in prevalence across racial and ethnic groups. The analyses outlined here provide details on the nature and extent of racial and ethnic health disparities in the *All of Us* participant cohort, and the accompanying browser can serve as a resource through which researchers can explore these disparities.

Introduction

The US National Institute on Minority Health and Health Disparities (NIMHD) defines health disparities as "health difference[s] that adversely affect disadvantaged populations" (1). The causes of health disparities are varied and span a multitude of biological, environmental, and social factors (2). A resource providing a means of identifying diseases for which there are health disparities as well as the groups that these disparities impact could expedite health disparity research and intervention.

The All of Us Research Program (*All of Us)* is an initiative that was launched by the US National Institutes of Health (NIH) in 2015 with the goal of accelerating research on personalized medicine and health equity (3). The program advances this goal through the collection of genetic, environmental, demographic, and healthcare data from volunteers from across the United States, with a special emphasis on recruiting participants from groups that have been historically underrepresented in biomedical research (4). The diverse participant body that has resulted from these efforts provide an opportunity to chart the landscape of health disparities in the United States (5,6).

We developed the US Health Disparities Browser to enable researchers to explore disease prevalence differences among *All of Us* participant self-identified racial and ethnic (SIRE) groups. The systematic identification of health disparities in the *All of Us* cohort is an important first step, which can enable subsequent studies aimed at characterizing their etiology. The US Health Disparities Browser is constructed from *All of Us* participant electronic health record (EHR) data collected from 208,268 participants and includes prevalence data for 1,755 diseases organized into 17 disease categories. Users

can search, browse, and visualize racial and ethnic health disparities for diseases of interest, with options to sort diseases by overall prevalence and differences in prevalence across SIRE groups.

Methods

Study Cohort

The web browser was developed from participant data made available through the *All of Us* Researcher Workbench, a cloud-based platform through which approved researchers can interface with and analyze *All of Us* participant data. The *All of Us* volunteer participant body is composed of US adults who enrolled in the program either electronically or through a partnered healthcare provider. All participants provided informed consent to participate in the program. The participant inclusion criteria include adults 18 and older, with the legal authority and decisional capacity to consent, and currently residing in the US or a territory of the US. Exclusion criteria exclude minors under the age of 18 and vulnerable populations (prisoners and individuals without the capacity to give consent). Details on participant recruitment, informed consent, inclusion and exclusion criteria are available online at https://allofus.nih.gov/sites/default/files/All_of_Us_Protocol_Overview_Mar_2022.pdf.

All of Us participant data available on the Researcher Workbench include answers to survey questions, lab measurements, electronic health record (EHR) data, and genomic data. The data is distributed across three access tiers that differ in sensitivity. The public tier dataset, which contains aggregate statistics, is freely accessible. The registered tier dataset consists of de-identified participant-level data and is restricted to registered researchers. The controlled tier dataset consists of participant-level genomic data. The browser relies on data from the All of Us Registered Tier Dataset v6 (curated version R2022Q2R2).

Race and Ethnicity

Participants enrolled in *All of Us* are asked to respond to surveys pertaining to details regarding their background and health. A survey titled “The Basics” asks participants “Which categories describe you? Select all that apply. Note: You may select more than one group”. The groups are (1) American Indian or Alaska Native, (2) Asian, (3) Black, African American, or African, (4) Hispanic, Latino, or Spanish, (5) Middle Eastern or North African, (6) Native Hawaiian or other Pacific Islander, and (7) White. Participants were also given the option to respond with “None of these fully describe me” or “Prefer not to answer”. Participant data on American Indian or Alaska Native identity are currently unavailable on the *All of Us* Researcher Workbench and thus were not included in this study.

As per the current US Office of Management and Budget Standards (OMB) for the classification of federal data on race and ethnicity, participant survey answers are coded as two variables in the Researcher Workbench: ‘race’ and ‘ethnicity’. The ‘ethnicity’ variable consists of information indicating whether a participant identified as Hispanic or Latino or not. The ‘race’ variable indicates any of the other seven categories the participant may have picked. The ‘race’ variable will list “More than one population” if more than one category other than Hispanic or Latino was selected. Consistent with these standards, we

defined Asian, Black, Middle Eastern or North African (MENA), Native Hawaiian or Pacific Islander (NHPI), and White participants as those who selected these respective racial categories in the “The Basics” core survey and no other racial or ethnic category. We defined Hispanic participants as all who selected “Hispanic, Latino, or Spanish”. We included an additional racial category in our browser titled “Multiple”, which consists of all individuals who selected two or more racial or ethnic categories other than “Hispanic, Latino, or Spanish”. These categories span a total of seven distinct *All of Us* participant SIRE categories used in this study.

Phenotype Case-Control Cohorts

All of Us participant diagnoses are coded in their EHR data as International Classification of Diseases codes (ICD-9-CM and ICD-10-CM). These codes were extracted and used to classify individuals as either disease cases or controls according to the phecode scheme outlined by the PheWAS consortium (7). The phecode scheme provides disease phenotype-specific inclusion and exclusion criteria from which case-control cohorts can be systematically created from ICD codes. A total of 1,755 unique phecode case-control cohorts were created from participant EHR data. Their corresponding phecodes belong to different phecode chapters spanning 17 disease categories. As we wished to control for sex, only participants who reported being assigned male or female at birth were included in calculations of disease prevalence estimates.

Quantifying Disease Prevalence

Female-born and older participants are largely overrepresented in the *All of Us* participant body. As such, overall and SIRE-specific disease prevalence estimates were adjusted for age and sex at birth. For overall and each SIRE group, unadjusted disease prevalence, *p*, was taken to be *K* cases over *n* total participants belonging to the group. Age and sex-adjustment was performed by weighing the unadjusted prevalence estimates of groups of participants corresponding to varying age-sex combinations using census fractions f. We define census fractions as the proportion of the total US population of a SIRE group g that falls into a particular age-sex group. Fractions were calculated from 2021 American Community Survey 1-year estimates. Adjusted prevalence values, $\hat{p}$, were calculated from different age-sex groups as:

$$\hat{p} = \sum_{g \in G} \frac{f_g}{n_g} K_g$$

95% confidence intervals for overall and SIRE-specific prevalence estimates were calculated by adding and subtracting the product of 1.96 and each adjusted estimate’s standard error, ($\hat{p}$), to and from the adjusted prevalence estimate:

$$\sigma\left(\hat{p}\right) = \sqrt{\sum_{g \in G} {f_g}^2 \frac{K_g}{n_g} \left(1 - \frac{K_g}{n_g}\right)}$$

$$p \epsilon [\hat{p} - 1.96\sigma(\hat{p}), \hat{p} + 1.96\sigma(\hat{p}\ )]$$

These confidence intervals are presented in the form of error bars in plots generated by the browser.

Quantifying Health Disparities

Racial health disparities in the *All of Us* participant body were quantified using three metrics: variance, range ratio, and range difference.

Variance was taken to be the average of the squared differences between seven SIRE-specific prevalence estimates p and average prevalence$\bar{p}$:

$$Variance = \frac{\sum \left(p_{Disease} - \bar{p}_{Disease}\right)^2}{7}$$

Range ratio was taken to be the binary logarithm of the ratio between the highest prevalence estimate observed among the seven SIRE-specific categories and the lowest prevalence estimate observed:

$$RangeRatio = log_2\left(\frac{Max\left(p_{Disease}\right)}{Min\left(p_{Disease}\right)}\right)$$

Range difference was taken to be the difference between the highest prevalence estimate observed among the seven SIRE-specific categories and the lowest prevalence estimate observed:

$$RangeDifference = Max\left(p_{Disease}\right) - Min\left(p_{Disease}\right)$$

The use of both range ratios and range differences allows for the capture of racial disparities in both diseases with low and high prevalence estimates, respectively.

Interactive Web Browser

Disease prevalence estimates and confidence intervals were calculated using version 1.21.6 of the numpy (8) and version 1.3.5 of the pandas (9) packages in Python 3. The plots displayed in the browser were generated using version 3.3.5 of the ggplot2 (10) package in R version 4.1.2. The web browser was created using version 1.7.4 of Shiny, a web application framework for R (11). The layout was designed using version 0.7.2 of the shinydashboard (12) package. Shiny Server was used to publish and host the application. The ssl certificate for the application was provided by Nginx.

Results

Participant Characteristics

Prior to participant exclusion via the phecode exclusion criteria, each phecode cohort consisted of 208,268 individuals (Table 1). Participants were predominantly White, assigned female at birth, and middle-aged or older. The cohort is racially and ethnically diverse; Black and Hispanic participants, in particular, are over-represented compared to their percentage of the US population.

Table 1
Characteristics of male and female born *All of Us* participants for whom EHR and SIRE data were available.

Characteristic	Count (%)
Complete cohort	208,268
Age	
20–29	13,333 (6.40)
30–39	29,628 (14.23)
40–49	28,852 (13.85)
50–59	36,488 (17.52)
60–69	46,320 (22.24)
70–80	37,979 (18.24)
80+	15,668 (7.52)
SIRE group	
Asian	5,410 (2.60)
Black	40,449 (19.42)
Hispanic	40,923 (19.65)
MENA	1,157 (0.56)
Multiple	3,223 (1.55)
NHPI	218 (0.10)
White	116,888 (56.12)
Sex	
Female	130,761 (62.78)
Male	77,507 (37.22)

The Landscape of Racial Health Disparities in All of Us

Racial and ethnic health disparities, as measured by range ratios and range differences in prevalence estimates across seven SIRE categories, can be found for a number of different conditions and diseases in the *All of Us* participant cohort. Diseases for which there are marked racial and ethnic prevalence disparities span a variety of different categories (Fig. 1).

Prevalence range difference between SIRE groups captures disparities among high prevalence diseases, whereas the range ratio better reflects disparities among low prevalence diseases. The conditions for

which the highest range differences were observed tended to be common metabolic conditions, such as hypertension, obesity, and hyperlipidemia, or disorders related to substance abuse, such as tobacco use disorder and other substance addictions. The conditions for which the highest range ratios were observed included sickle cell anemia, which disproportionately affects Black participants, and a number of dermatological diseases that affect White participants, such as actinic keratosis and skin cancer (Table 2).

Table 2
Top 10 health disparity diseases with the highest values for range difference and range ratio.

Disease	Prevalence (%)	Value
Range Difference		
Tobacco use disorder	18.92	26.75
Essential hypertension	45.50	25.92
Obesity	27.64	24.60
Hypertension	34.57	24.06
Morbid obesity	14.90	23.98
Substance addiction and disorders	14.30	22.58
Benign neoplasm of skin	15.79	21.50
Major depressive disorder	31.37	20.17
Anxiety disorder	30.94	18.49
Hyperlipidemia	40.21	18.08
Range Ratio		
Sickle cell anemia	1.02	6.33
Actinic keratosis	7.93	5.50
Hemangioma of skin and subcutaneous tissue	4.70	5.26
Chronic dermatitis due to solar radiation	5.27	5.03
Carcinoma in situ of skin	1.28	4.68
Melanomas of skin	1.36	4.61
Nevus, non-neoplastic	5.30	4.35
Other non-epithelial cancer of skin	3.35	4.26
Lyme disease	0.90	4.12
Squamous cell carcinoma	1.79	4.00

We also identified conditions and diseases that showed the highest group-specific prevalence for each SIRE group. Among these diseases high prevalent diseases, we identified the five that were the most disparate across the SIRE groups, as determined by variance in prevalence estimates (Table 3).

Table 3
High prevalence health disparity conditions and diseases for each SIRE group.

Disease	Group-specific Prevalence (%±CI)	Variance
Asian		
Atopic/contact dermatitis due to other or unspecified cause	19.36 ± 1.21	38.18
Myopia	11.35 ± 0.92	10.99
Disorders of vitreous body	10.56 ± 0.95	10.06
Acne	9.06 ± 0.83	8.63
Astigmatism	8.51 ± 0.81	7.29
Black		
Substance addiction and disorders	23.73 ± 0.54	46.40
Overweight, obesity and other hyperalimentation	18.22 ± 0.50	15.76
Hypertension	31.76 ± 0.71	15.63
Alcohol-related disorders	11.41 ± 0.42	13.87
Hypertensive chronic kidney disease	13.10 ± 0.51	13.37
Hispanic		
Viral infection	13.59 ± 0.44	18.25
Other conditions or status of the mother complicating pregnancy, childbirth, or the puerperium	9.78 ± 0.33	7.80
Gastritis and duodenitis, NOS	8.72 ± 0.33	6.59
Other complications of pregnancy NEC	8.92 ± 0.31	6.25
Dermatophytosis	7.45 ± 0.31	6.23
MENA		
Vitamin D deficiency	22.99 ± 2.89	44.27
Diseases of esophagus	19.22 ± 2.74	34.02
Esophagitis, GERD and related diseases	18.95 ± 2.72	33.73
Other disorders of bone and cartilage	17.10 ± 2.45	25.83
Spondylosis without myelopathy	16.83 ± 2.40	24.45

Disease	Group-specific Prevalence (%±CI)	Variance
Asian		
Multiple		
Bipolar	12.40 ± 1.62	18.77
Asthma	22.26 ± 1.72	14.22
Posttraumatic stress disorder	10.66 ± 1.54	12.81
Asthma with exacerbation	8.98 ± 1.18	8.83
Attention deficit hyperactivity disorder	6.06 ± 0.83	4.32
NHPI		
Tobacco use disorder	20.01 ± 7.61	62.54
Morbid obesity	12.99 ± 7.21	38.03
Obesity	24.26 ± 7.75	36.87
Essential hypertension	36.51 ± 9.38	28.77
Type 2 diabetes	18.46 ± 7.49	18.91
White		
Allergic rhinitis	19.06 ± 0.31	44.94
Benign neoplasm of skin	12.77 ± 0.27	39.39
Anxiety, phobic and dissociative disorders	16.44 ± 0.32	39.08
Anxiety disorder	29.59 ± 0.38	35.79
Major depressive disorder	29.52 ± 0.37	33.65

Interactive Web Browser

We developed the US Health Disparities Browser to catalog and present the landscape of health disparities in the *All of Us* participant cohort. The browser and its underlying database enable users to assess how prevalence estimates for specific diseases and phecodes of interest differ across seven SIRE groups in the *All of Us* participant cohort (Fig. 2). Disease prevalence values and confidence intervals are presented as a bar plot with different bars representing different SIRE groups. Overall case and control counts and adjusted prevalence are displayed above the generated bar plots. The browser also features an interactive catalog of the 1,755 diseases featured in the browser, presented in table form, which can be

sorted by disease name, phecode, overall prevalence across the *All of Us* participant body, and variance across SIRE groups. Users can search the disease disparity catalog by disease name or phecode.

The results reported on the browser comply with the *All of Us* Data and Statistics Dissemination Policy. The browser shows summary statistics only, does not reveal participant-level data in any way, and does not display any participant group count ≤ 20. The browser is not affiliated with, nor endorsed by, the *All of Us* Research Program, the NIH, or the US Department of Health & Human Services.

Discussion

Through the creation of the US Health Disparities Browser and its accompanying analyses, we report marked disparities in disease prevalence estimates across racial and ethnic groups in the *All of Us* cohort. The most profound disparities that were observed appear to be concentrated in metabolic and dermatological conditions, as well as ailments related to substance use.

The analyses outlined here are limited by sampling bias, as the *All of Us* participant body is composed of volunteer participants. Furthermore, these analyses rely on EHR data, which may reduce their utility in drawing insight from non-chronic diseases and from populations with inadequate access to healthcare.

Nonetheless, the results described here provide evidence for racial and ethnic health disparities in the *All of Us* participant body, consistent with what is known for the US population (13–15). The interactive web browser that was developed alongside this study could serve as a useful resource for researchers interested in leveraging the *All of Us* platform to conduct health disparities research, since the identification of disparity diseases is a prerequisite for subsequent etiological studies. The potential of this resource is amplified by the rich genetic, environmental, and demographic data being collected for the large and diverse *All of Us* participant cohort.

Declarations

Acknowledgements

VL and LMR were supported by the Division of Intramural Research (DIR) of the National Institute on Minority Health and Health Disparities (NIMHD) at NIH, (Award Numbers: 1ZIAMD000016 and 1ZIAMD000018). LMR was supported by the National Institutes of Health (NIH) Distinguished Scholars Program (DSP). SS and IKJ were supported by the by the IHRC-Georgia Tech Applied Bioinformatics Laboratory (Award Number: RF383).

Conflict of interest

The authors declare no conflict of interest.

References

1. NIMHD (2022) Minority Health and Health Disparities: Definitions and Parameters. U.S. Department of Health and Human Services, Bethesda, MD.
2. Adler, N.E., Rehkopf, D.H. (2008) U.S. disparities in health: descriptions, causes, and mechanisms. *Annu Rev Public Health*, **29**, 235-252.
3. All of Us Research Program, I., Denny, J.C., Rutter, J.L.*, et al.* (2019) The "All of Us" Research Program. *N Engl J Med*, **381**, 668-676.
4. Mapes, B.M., Foster, C.S., Kusnoor, S.V.*, et al.* (2020) Diversity and inclusion for the All of Us research program: A scoping review. *PLoS One*, **15**, e0234962.
5. Ramirez, A.H., Sulieman, L., Schlueter, D.J.*, et al.* (2022) The All of Us Research Program: Data quality, utility, and diversity. *Patterns (N Y)*, **3**, 100570.
6. Kathiresan, N., Cho, S.M.J., Bhattacharya, R.*, et al.* (2023) Representation of Race and Ethnicity in the Contemporary US Health Cohort All of Us Research Program. *JAMA Cardiol*, **8**, 859-864.
7. Bastarache, L. (2021) Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci*, **4**, 1-19.
8. Harris, C.R., Millman, K.J., van der Walt, S.J.*, et al.* (2020) Array programming with NumPy. *Nature*, **585**, 357-362.
9. The pandas development team (2020) pandas-dev/pandas: Pandas. *Zenodo*.
10. Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, New York, NY.
11. Winston Chang, J.C., JJ Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, Barbara Borges (2023) shiny: Web Application Framework for R.
12. Winston Chang, B.B.R. (2021) shinydashboard: Create Dashboards with 'Shiny'.
13. National Center for Health Statistics (2016) Health, United States, 2015: With special feature on racial and ethnic health disparities.
14. LaVeist, T.A. (2011) *Minority populations and health: An introduction to health disparities in the United States*. John Wiley & Sons.
15. Baciu, A., Negussie, Y., Geller, A.*, et al.* (2017) The state of health disparities in the United States. *Communities in action: Pathways to health equity*. National Academies Press (US).

Figures

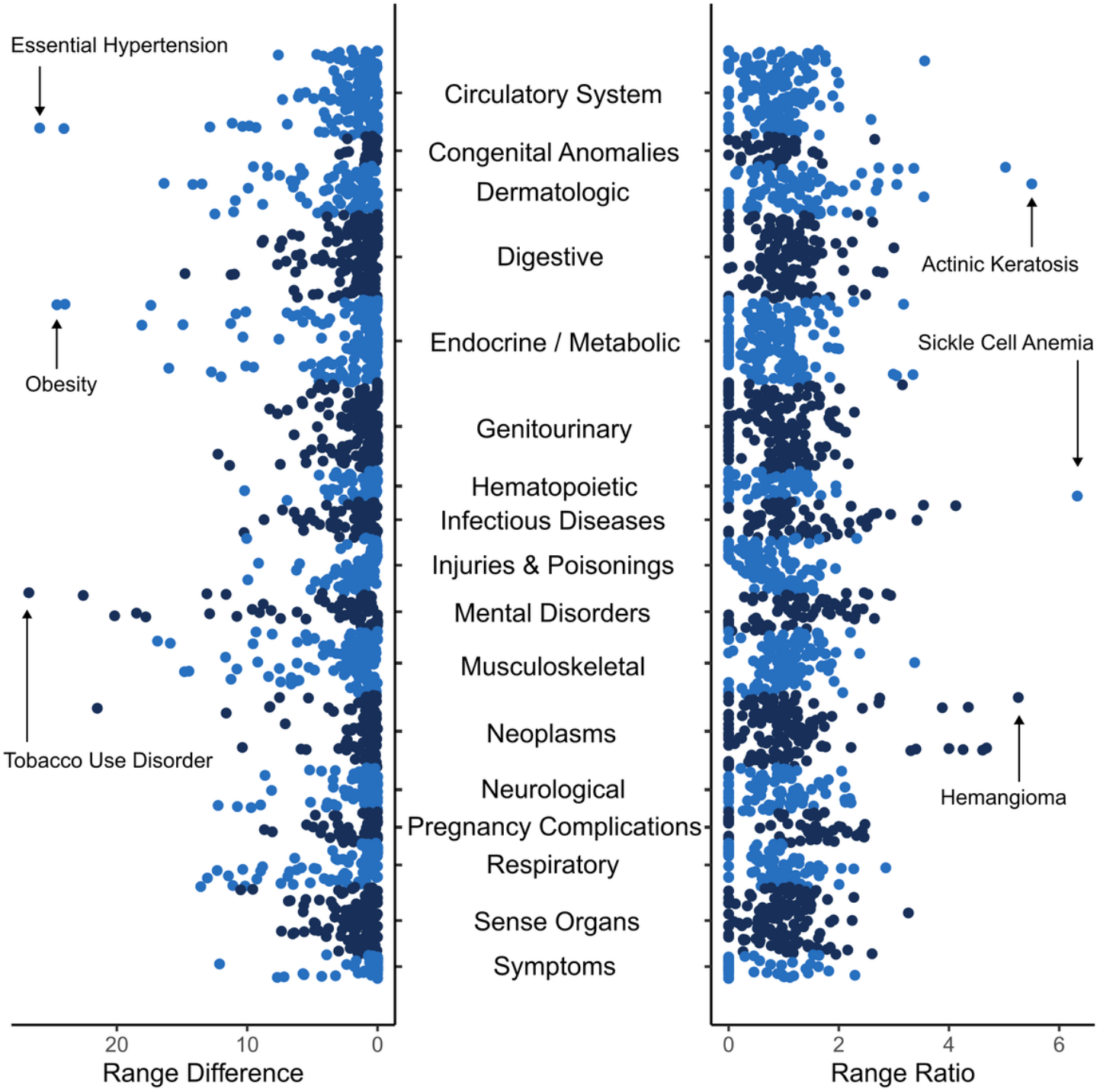

Figure 1

Racial and ethnic disparities in disease prevalence estimates. Disease prevalence disparities are quantified by prevalence range difference and prevalence range ratio among SIRE groups as described in the methods. Each point represents a distinct disease or condition, organized by disease categories.

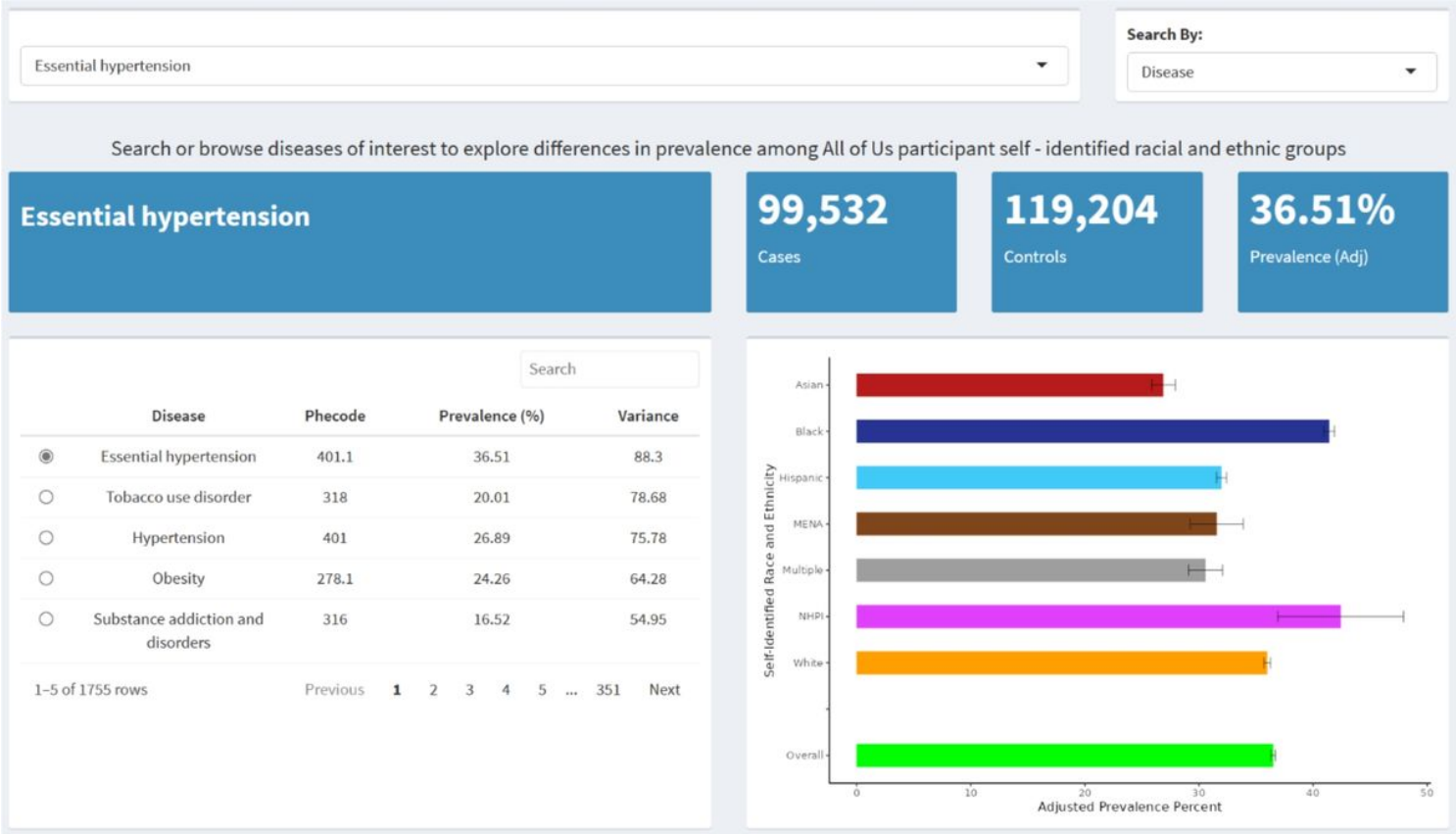

Figure 2

The US Health Disparities Browser. Screenshots of the browser showing basic functions and information provided upon querying a disease or phecode.