ARTICLE

Infrastructure

Leveraging the All of Us Database for Primary Care Research with Large Datasets

Daniel J. Parente, MD, PhD

The National Institutes of Health (NIH) are supporting the *All of Us* research program, a large multicenter initiative to accelerate precision medicine. The *All of Us* database contains information on greater than 400,000 individuals spanning thousands of medical conditions, drug exposure types, and laboratory test types. These data can be correlated with genomic information and with survey data on social and environmental factors which influence health. A core principle of the *All of Us* program is that participants should reflect the diversity present in the United States population.

The *All of Us* database has advanced many areas of medicine but is currently underutilized by primary care and public health researchers. In this Special Communication article, I seek to reduce the "barrier to entry" for primary care researchers to develop new projects within the *All of Us* Researcher Workbench. This Special Communication discusses (1) obtaining access to the database, (2) using the database securely and responsibly, (3) the key design concepts of the Researcher Workbench, and (4) details of data set extraction and analysis in the cloud computing environment. Fully documented, tutorial R statistical programming language and Python programs are provided alongside this article, which researchers may freely adapt under the open-source MIT license. The primary care research community should use the *All of Us* database to accelerate innovation in primary care research, make epidemiologic discoveries, promote community health, and further the infrastructure-building strategic priority of the family medicine 2024 to 2030 National Research Strategy. (J Am Board Fam Med 2024;37:S144–S155.)

Keywords: **ADFM/NAPCRG Research Summitt 2023, All of Us Database, Cloud Computing, Database, Multicenter Studies, Family Medicine, Precision Medicine, Primary Health Care, Public Health, Retrospective Studies, Statistics**

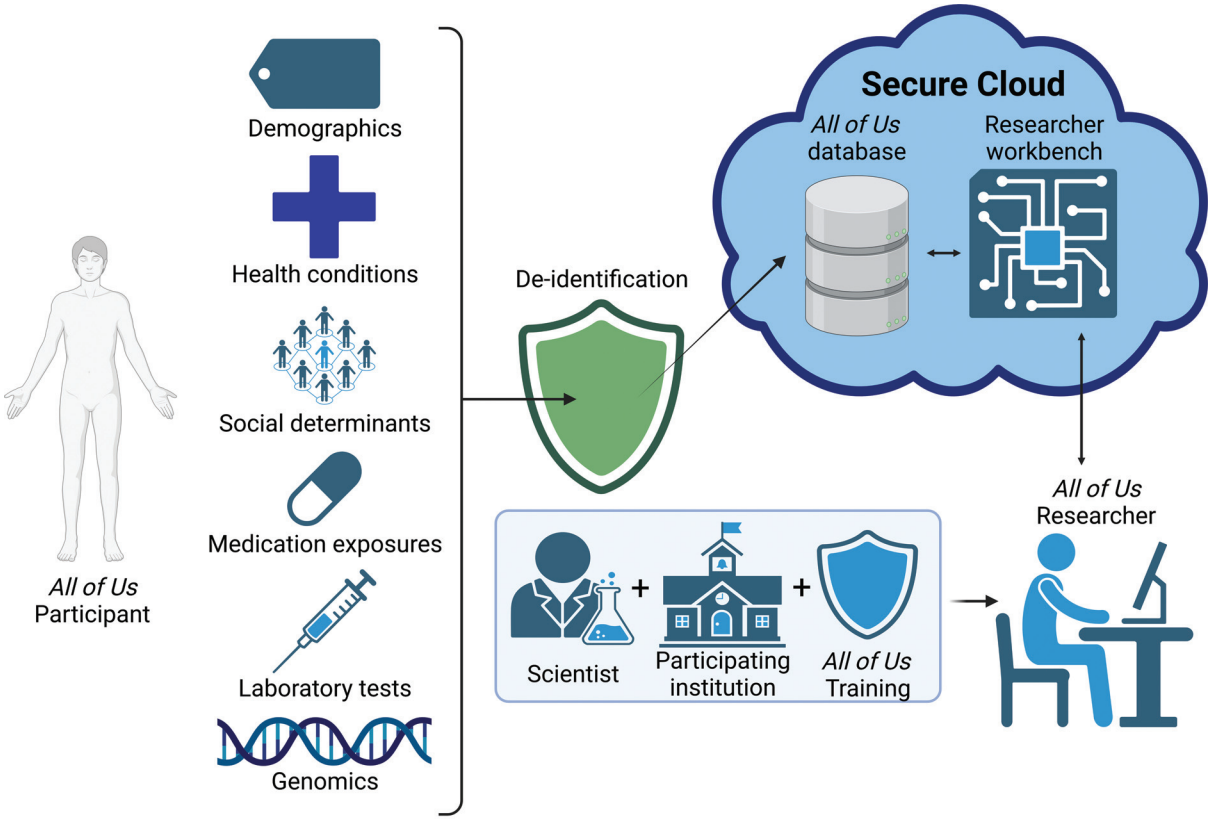
The *All of Us* Research Program and Primary Care Research

The *All of Us* research program is a large, multicenter initiative backed by the National Institutes of Health (NIH) to accelerate precision medicine.[1] The *All of Us* research program has a target of recruiting 1 million or more participants to share data derived from electronic health records, biospecimens, and surveys in a centralized database. It also has as a core value that the set of participants "should reflect the broad diversity present in the United States population."[2] In its current form, it contains information for greater than 400,000 individuals spanning more than 25,000 distinct medical conditions, 29,000 drug exposure types, and 16,000 laboratory test types; 245,000 participants have small nucleotide polymorphism (SNP) and insertion/deletion (indel) variants annotated from short-read whole genomic sequencing, and an additional 1,000 participants have long-read whole genomic sequencing data available. *All of Us* survey data captures demographic and environmental factors variables that influence health, which allows these factors to be considered alongside biological variations to understand the development and trajectory of human disease more fully. (Figure 1)

The *All of Us* database has been used to make advances across broad areas of medicine including

This article was externally peer reviewed.
Submitted 6 December 2023; revised 5 March 2024, 29 March 2024; accepted 1 April 2024.
From the Family Medicine and Community Health, University of Kansas Medical Center, Kansas City, KS (DJP).
Funding: This work was funded by a contract from the American Academy of Family Physicians.
Conflict of interest: None.
Corresponding author: Daniel J. Parente, MD, PhD, University of Kansas Medical Center, 3901 Rainbow Boulevard, MS 4010, Kansas City, Kansas 66160 (E-mail: dparente@kumc.edu).

Figure 1. Schematic diagram of the *All of Us* research platform.



Note: All of Us participants are recruited through a national program. Demographic, medical, and social data are collected, deidentified and stored securely in a cloud-based environment. Scientists from participating institutions who have completed All of Us training ("All of Us researchers") can perform analyses on the data using a cloud computing strategy through the "Researcher Workbench." Graphics were created with BioRender.

cancer[3–13]; infectious disease (including COVID-19)[14–23]; pulmonary medicine[10,24]; cardiovascular medicine[6,25–34]; obesity[35–37]; endocrine, diabetes and thyroid disorders[31,34,38–44]; neurological, stroke, and neurocognitive disorders[31,45–47]; dermatologic, breast and integumentary disorders[6,8,16,21,24,42,45,48–76]; transplant medicine[50]; rheumatology, allergies and autoimmune medicine[24,51,56,59,61,73,76,77]; substance use disorders[52,57,78,79]; genomic medicine[11,12,33,80]; ophthalmology[29,39,40,43,72,78,79,81–86]; and gastroenterology and hepatology.[13,17,56,87]

Research using the *All of Us* database is in many cases directly relevant to the practice of primary care. Here I highlight 4 recent examples. First, recent work determined that the prevalence of hypertension in the *All of Us* cohort was 27.9%, and that this rate was similar to that previously observed in the National Health and Nutrition Examination Survey (NHANES) cohort.[27] Epidemiologic studies of this nature are important to understand disease burden and influence policy about how resources should be allocated toward primary care. Second, "real-world" use of sodium glucose cotransporter 2 inhibitors (SGLT2i) and glucagon-like peptide 1 receptor agonists (GLP1-RA) were evaluated. Use of these medications was shown to be low, particularly among under-represented groups.[41] In a third study, hypertension and type 2 diabetes were shown to be major risk factors for dementia.[31] Multi-variable analysis also suggested that the interaction of race and ethnicity with hypertension indicates a need for targeted interventions in primary care settings in at-risk, structurally disadvantaged populations.[31] Fourth, in the last study I highlight here, greater physical activity – as measured by step counts using Fitbit wearable devices – was associated with decreased risk of type 2 diabetes, and that multivariable analysis indicated that there was no significant modification of this effect by age, sex, body mass index, or sedentary time.[44]

Table 1. Glossary of Terms

Term	Meaning	Examples
Concept	A piece of information about a participant. Concepts may include both the information itself (e.g., a survey response) and metadata about the information (e.g., time of collection, survey version).	- A participant's age at time of consent - A survey question about experience of food insecurity
Concept set	A group of related concepts.	- The social determinants survey asks questions about whether a respondent experienced food insecurity, or only worried about experiencing food insecurity. A concept set containing both questions would represent a response to either (or both) of these questions.
Value	A specific data element (a number, a text string, etc.) from a concept.	- The participant's answer to a survey question - The time and date the participant completed a survey - The version of the survey administered
Cohort	A population of interest, built using Boolean logic ("AND" and "OR" relationships) between concepts and concept sets.	- All participants who had both a Vitamin B12 measurement AND answered at least one of the two food insecurity questions on the social determinants of health survey. - All adults with diagnosis of diabetes
Dataset	A set of tables extracted from the *All of Us* database, chosen by specifying concept sets, and then selecting values from the concept sets to extract.	- A demographics table containing the participant's code number and age, plus a measurements table containing a participant's code number, Vitamin B12 serum concentration, and the time the laboratory test was collected.

I aim to highlight the promise of the NIH *All of Us* database for advancing primary care research. Concordant with the National Research Strategy that was recently developed at the 2023 National Family Medicine Research Summit,[88] I encourage the primary care research community to use the *All of Us* database to enhance the practice of family and community medicine. To that end, I highlight detailed methods as a "tutorial walkthrough" to implementing an example project using the *All of Us* database. As an example, I explored the relationship between food insecurity and Vitamin B12 deficiency. The background, results, and clinical meaning of this association are discussed separately in a companion article. Here, I focus on replicability of the methods so that the barrier to use of the *All of Us* database is reduced for other primary care researchers.

The Researcher Workbench

All of Us makes data available to researchers through registration with the Researcher Workbench, available at https://www.researchallofus.org/data-tools/workbench/ (Figure 1). Researchers may gain access if they meet 2 conditions. First, they must complete training through the *All of Us* portal on issues related to data security and appropriate use of research data. A particular focus of this training is avoiding exacerbating pre-existing structural inequalities in health care. Completion of this training grants access to the "registered tier." Additional training is required to access data within the "controlled tier," which includes genomic data. Second, researchers must be members of an institution that has a data use and registration agreement with *All of Us*. This currently includes more than 600 participating institutions; an up-to-date list is available on their website at https://www.researchallofus.org/institutional-agreements/.

To create a project, researchers are required to publicly report the primary purpose of their project, which may be for research, educational, for-profit, or other purposes. Researchers must specify whether the project is: disease-focused research; methods development and validation; to develop controls, such as for comparisons with another data set from a different resource; social and behavioral research; population health or public health research; ethical, legal, and social implications research; or drug and therapeutics development research. The purpose of this project will be publicly displayed. Likewise, users must provide a publicly available summary of their research purpose by describing the scientific questions that are being evaluated by the study, and describing the relevance to science and public health. The scientific approach must also be briefly explained, and anticipated findings from the study must also be

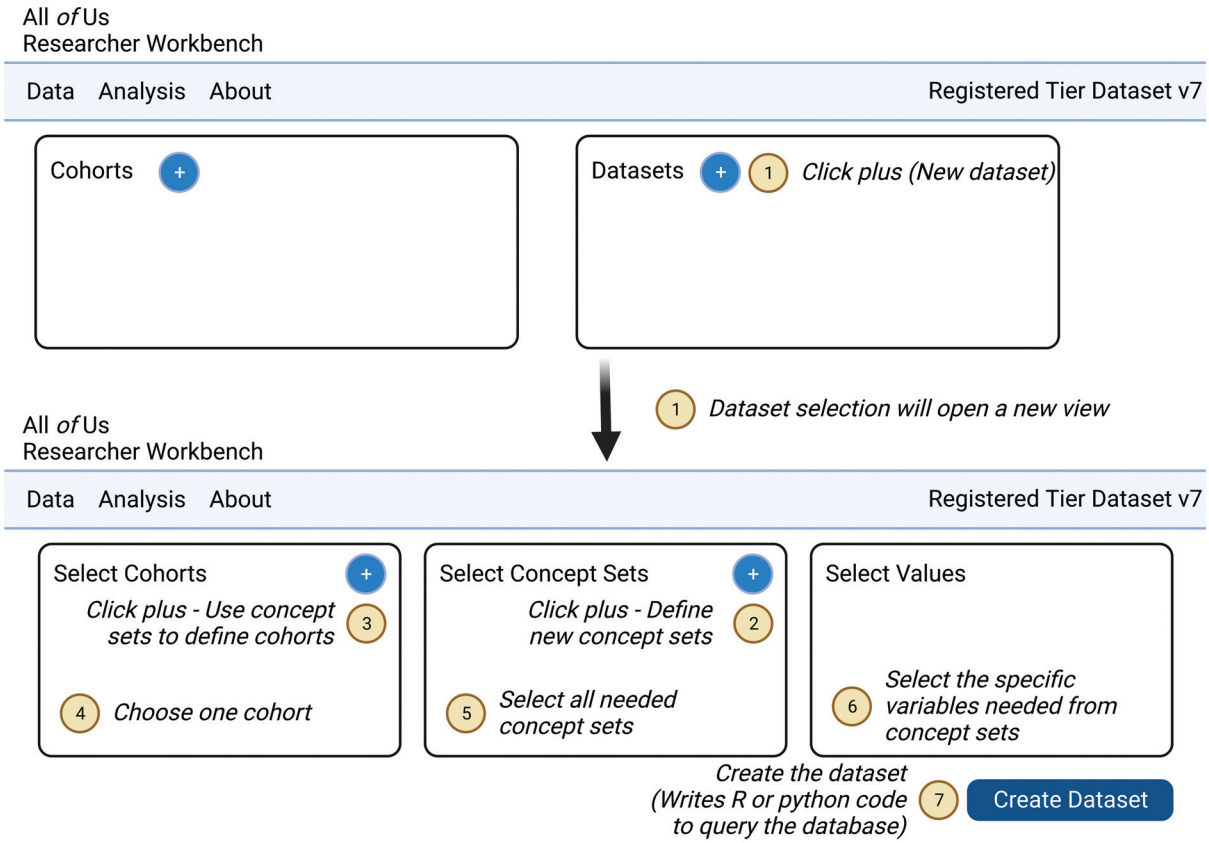
expressed. Researchers must report the population of interest to be studied, and if the population is a historically underrepresented population, this must be specified.

The Workbench uses "concepts," "concept sets," "values," "cohorts," and "datasets." (Table 1). These are used to specify the population of interest and extract relevant variables from the database, as discussed in the next sections.

Concept Sets

"Concept sets" can be defined to group related concepts together. This promotes ease of extraction of these variables when building datasets, and for ease of building cohorts (see next section). In this example, I define 3 concept sets: a vitamin B12 concept set, a food insecurity concept set, and a metformin-exposure concept set. The food insecurity concept set comprises 2 concept corresponding to whether the participant expressed (1) an experience of or (2) worry about food insecurity. The *All of Us* Researcher Workbench's design implicitly suggests that users should first define a cohort, second define concept sets, and third define values for export. Because cohorts depend on data elements that are in concept sets, I instead suggest that users first define concept sets before defining a cohort (see Figure 2, which demonstrates the recommended workflow). Once a concept set is defined, it can then be readily used to define a cohort. Concept sets are therefore central to both cohort definition and dataset extraction.

Figure 2. Recommended workflow for defining concept sets, cohorts, and datasets.



Note: A schematic of the recommended workflow is shown. The upper panel shows the main screen presented to researchers in the All of Us Researcher Workbench. Related variables are grouped into "concept sets," which are then used for 2 purposes. First, they can be used to define a cohort. Second, they can be used to select data elements for extraction. Because of the centrality of cohort sets in both these processes, I recommend defining them first using the workflow (yellowed number circles) provided below in the following order: first, begin creating a new dataset; second, define concept sets; third, use concept sets to define cohorts; fourth, select a cohort; fifth, select concept sets for extraction; and sixth, select the specific variables for extraction, derived from the concept sets. Graphics were created with BioRender.

Cohorts

Cohorts can be defined using a visual builder implementing Boolean logic ("AND" and "OR" relationships) between concepts and concept sets to define a population of interest. In my example about food insecurity and Vitamin B12 deficiency, I created a cohort of all participants who had both a Vitamin B12 measurement AND answered at least 1 of the 2 food insecurity questions on the social determinants of health survey. Specification of these constraints amounts to an AND condition between the Vitamin B12 concept set and the Food Insecurity concept set. For greater plausibility in the relationship between survey response and laboratory measurement, I wish to constrain these measurements to have occurred around the same time (eg, within a year). In principle, the visual interface in *All of Us* allows for expression of this sort of temporal constraint. In my hands, however, I observed that attempts to specify temporal queries often led to errors in the interface. I found it much easier to enforce temporal relationships between concepts during analysis in the cloud computing environment after exporting the dataset.

Dataset Extraction

Once concept sets and a cohort have been defined, a "dataset" can be constructed. A dataset is a group of concept sets applied to a cohort and requires choosing specific values to extract. The *All of Us* database contains a group of common concept sets that many projects may use, for example, demographics, and also allows users to define custom concept sets, as described above. For the Vitamin B12 and food insecurity example investigation, I used the above-discussed cohort and, from this, extracted (1) demographics, (2) vitamin B12 levels, (3) food insecurity-related answers to the social determinants survey, and (4) data on drug exposures (metformin). Note that concepts can be extracted even if they are not part of the cohort definition; in this example, presence/absence of metformin exposure was not part of the cohort definition because I want to capture the experience of individuals both with and without metformin exposure.

To extract the dataset *All of Us* writes a set of custom scripts in either R[89] or Python[90] (user selectable) containing a SQL database query to extract all specified values, and a few lines that save the results to the persistent cloud storage environment. These scripts can be run in the cloud computing environment, described below, as the first part of an analysis pipeline.

In this example, data are extracted from the database in 4 tables: person (demographics), measurements (Vitamin B12 levels), survey (answers to the food insecurity questions), and drug (metformin exposure). Concrete examples of integrating data from these tables is discussed in the accompanying example scripts available at https://github.com/djparente/AllOfUsExample and in Supplemental Materials.

Cloud Computing Environment

Data analysis is performed in a secure environment within the *All of Us* Researcher Workbench. Costs for using this environment are low, but not entirely free in the long-term, and depend on the computing resources required to conduct the project. The default environment recommended for general analysis by *All of Us* is a 4-CPU system with 15 GB of RAM, without a GPU. While the environment is running there is a cost of $0.20 per hour. This environment can be paused with a cost of "<$0.01" per hour. Persistent disk storage, at time of writing, cost $4.80 per month. Thus, for a project worked on for 80 computer hours over 6 months, a total computer cost of approximately $44.80 ($0.20/h x 80 hours + $4.80/mo x 6 months) would be incurred.

Use of fewer resources can reduce computer-time costs. For example, a 1-CPU system with 3.75 GB of RAM incurs an hourly cost of only $0.06 per hour. However, total costs are likely to be primarily driven by the duration storage is required (in this example, 6 months) rather than the costs of the computing environment. Using the least-resourceful system in the above example, the cost would be reduced only to $33.60. For my analyses on thousands of participants using descriptive techniques and logistic regression, I found the default configuration to be more than adequate. Greater computing resources are available for much higher costs per hour (eg, 96-CPU system with hundreds of GB of RAM and multiple graphical processing units [GPUs]). Conceivably such resource-intensive configurations might become important for very advanced applications, such as those driven by artificial intelligence and machine learning based on transformer architectures[91] or local large language models;[92–94] or for advanced genomics research.

More complex projects involving transferring exceptionally large volumes of database data (gigabytes or terabytes), or some workflows involving genomics, might incur additional costs. I suspect, however, that most primary care researchers will find the low-cost, default environment adequate for most of their purposes.

To defray initial start-up expenses, the *All of Us* researcher workbench provides $300 in free computing credits, which allows exploration of the research environment. Users can configure the computing environment to automatically suspend itself when idle for a duration (by default, 30 minutes) to prevent unintentional overuse of computing resources. This was completely adequate for my purposes; my initial projects familiarizing myself with the interface incurred <$30 in expenses, and I expect this experience to be true of most primary care researchers.

Data Preparation and Statistical Analysis

Within the workbench, researchers can use the R statistical computing language or the Python programming language to analyze data and perform statistical calculations. The default computing environments comes preloaded with many of the most used statistical and data operations packages for R and Python, but researchers are not limited to them. More packages can be installed, as needed, using the Python package index (PyPI) via pip, and through R's "install.packages()" command. When a new computer environment is created, these packages may need to be reinstalled, so I recommend placing installation commands at the beginning of *All of Us* analysis scripts. The analysis environment is thus very flexible and can be readily adapted to many common workflows.

For the Vitamin B12 and food insecurity example, I have developed a set of example scripts that highlight data preparation in Python and statistical analysis in R. These example scripts are available at https://github.com/djparente/AllOfUsExample and in Supplemental Materials. These example applications demonstrate many common situations that researchers will likely encounter including, (1) transformation of survey data into usable formats, (2) extraction of numeric laboratory data, (3) extraction of drug exposure information, (4) enforcement of temporal constraints on the dataset (eg, considering only Vitamin B12 levels within 12 months of a survey response), (5) merging data together from various data tables (demographics, drug exposures, laboratory data, survey responses), (6) exclusion of outliers, (7) dichotomizing ordinal variables, and (8) conducting logistic regression and controlling for confounding variables. These example scripts may be freely adapted under the open-source MIT license for use in other *All of Us* projects. They also demonstrate how sophisticated analysis can be accomplished using only a small amount of code in the *All of Us* Researcher Workbench. In addition, because the example scripts use Python to do data preparation and R to do statistical analysis, it demonstrates how both frameworks can be used together and can exchange data through persistent disk cloud storage. If use of only 1 framework is desired, then data preparation and statistical analysis can be performed in either pure Python or pure R.

Users who are familiar with R and/or Python will find adapting to the cloud computing environment straightforward. Those familiar with SAS, SPSS or Stata will unfortunately not yet find these tools available in the *All of Us* cloud environment, but can find analogous statistical analysis functions in R. The addition of SAS, SPSS and Stata support to the *All of Us* environment in the future would further reduce barriers to use of the environment by researchers who might prefer these tools. Clinicians who have less experience with statistical analysis, R or Python may find collaboration with a data scientist or biostatistician to be mutually beneficial to all involved. Data scientists may benefit from clinician guidance toward questions likely to impact real-world clinical practice. Likewise, clinicians may benefit from data scientist/biostatistical expertise to ensure that analyses arrive only at robust, rigorous, and valid conclusions.

Human Subjects Protection

Collection of participant data are overseen by the NIH *All of Us* Institutional Review Board (IRB). The data portal has undertaken significant efforts to deidentify data and to reduce the risk of incidental reidentification. Data at the "Registered" access tier been (1) stripped of explicit identifiers and geolocation data, (2) undergone random shifting of dates, and (3) had removal of data from participants more than the age of 89 years. I expect that most analyses at the "Registered" access tier will be considered by local IRBs as "Not Human Subjects Research" as they are secondary data analyses on deidentified datasets. This example project was reviewed by the University of Kansas Medical

Center IRB and determined to be Not Human Subjects Research. Nonhuman subjects research status may allow primary care researchers to analyze data rapidly with fewer regulatory barriers. By contrast, data in the "Controlled" access tier – including whole genomic sequencing, genotyping arrays, unshifted dates, and some suppressed demographic features – may more likely be considered "human subjects research" because those data elements are more likely to be considered by an IRB as "identifiable private information" under the 45 CFR 46[95] definition of human subjects research. These considerations notwithstanding, I advise researchers to seek guidance from their local IRB on the protection of human subjects before initiating any *All of Us*-based projects.

Dissemination of Results

The *All of Us* research program requires researchers to notify the program at least 2 weeks before publications or presentations resulting from analyses of the data. Publications do not require approval from *All of Us*, but this notification is required to allow the program to prepare for media coverage or other issues that may be related to *All of Us* data. In addition, a final peer-reviewed manuscript must be deposited within PubMed Central immediately on acceptance for publication.

Primary Care Researchers Should Use All of Us to Accelerate Precision Medicine Research

In summary, the *All of Us* research program is an ambitious, multi-center effort to accelerate precision medicine. It has been used to advance many diverse areas within medicine.[3–87,96] Here, I highlight the potential for application of the *All of Us* database to address important issues within primary care, community medicine, and public health. I provide a detailed methodological template of using the database through the example of testing a hypothesis about Vitamin B12 deficiency and food insecurity. I hope this example and detailed instructions will reduce the technical "barrier to entry" to use of the *All of Us* platform, thereby driving greater primary care researcher engagement with it.

Leaders in family medicine recently met for the 2023 National Family Medicine Research Summit to create a national strategy for family medicine research between 2024 and 2030.[88] This plan calls for research that is "whole-person, family, and community centered and improves health by enhancing health promotion, improving care for chronic diseases and advancing health care delivery, whereas including cross-cutting themes of health equity, technology, and team science."[88] Among the 3 strategic priorities identified by this plan is the development of a "national infrastructure for organizing and optimize family medicine research opportunities."[88] Within this strategic plan, primary care researcher use of the *All of Us* database can help accomplish some of the specific infrastructure-related objectives conceived of by the plan, viz.:

- C2. Utilize a repository of clinical data to answer key questions in primary care, and
- C5. Design and utilize distinctive methodology such as . . . *big data analytics* and machine learning (emphasis mine)

Primary care researchers will find that *All of Us* data can be used to improve the health of their patients and communities by uncovering associations between disease processes and social, environmental, and biological determinants of health. The *All of Us* database holds the potential to uncover new insights into disease processes that primary care clinicians frequently encounter, including obesity, type 2 diabetes, cardiovascular disease, cerebrovascular disease, and cancer. Primary care clinicians may also find benefit in research conducted by specialists on topics that are related to primary care. For example, there is a large and growing body of research works that use *All of Us* data to better understand dermatologic disease.[6,8,16,21,24,42,45,48–76] Primary care clinicians are often the "front line" clinicians who first encounter these disease processes. I nevertheless propose – and in concordance with the vision of the family medicine National Research Strategy[88] – that primary care clinicians should not be content to merely observe specialist-conducted research on the *All of Us* database. Primary care involves implementing plans for *whole persons*, which is not adequately informed by the sum of research on individual organ systems. Primary care researchers must take a leadership role in using the *All of Us* database to enhance primary care and community health.

Many problems in primary care are strongly influenced by the social and environmental context of patients. Greater understanding of the associations between these factors and human disease will allow primary care clinicians to better care for whole persons. In doing so, they may

also gain insight that will help alleviate pre-existing structural inequalities in health care that result from ongoing and historic biases. I invite the primary care research community to use *All of Us* to better understand these issues.

One future application for the *All of Us* database may be to accelerate research activities among family medicine resident physicians. Resident physicians must complete scholarly activities as part of ACGME requirements.[97] I envision the construction of a "resident research pipeline" in which (1) resident physicians develop clinical questions inspired by their unique experiences during residency, (2) perform a literature search on what is already known about their questions, (3) construct a relevant cohort and data set in the *All of Us* portal, and (4) are guided in using bivariable and multivariable methods to answer their research question. Developing such a pipeline would allow residents to conduct potentially high-impact, multicenter research with results likely to generalize to other settings.

In closing, I call on the primary care research community to make use of the NIH *All of Us* database to accelerate innovation, make epidemiologic discoveries, and promote community health. Maximizing the generalizability and impact of primary care research conducted using the database may benefit from the formation of multi-institutional, national teams to develop and implement inquiries on topics of high clinical relevance to primary care. Moreover, I call on organizations dedicated to primary care – including the American Board of Family Medicine, the American Academy of Family Physicians, the Society of Teachers of Family Medicine, the Association of Departments of Family Medicine, and the North American Primary Care Research Group – to actively promote and devote resources to using the *All of Us* database to enhance the quality of primary care.

To see this article online, please go to: http://jabfm.org/content/37/S2/S144.full.

References

1. Denny JC, Rutter JL, Goldstein DB, All of Us Research Program I, et al. The "All of Us" research program. N Engl J Med 2019;381:668–76.
2. Mapes BM, Foster CS, Kusnoor SV, All of Us Research Program, et al. Diversity and inclusion for the All of Us research program: a scoping review. PLoS One 2020;15:e0234962.
3. Aschebrook-Kilfoy B, Zakin P, Craver A, All of Us Research Program Investigators, et al. An overview of cancer in the first 315,000 All of Us participants. PLoS One 2022;17:e0272522.
4. Coughlin SS, Chen J, Cortes JE. Health care access and utilization among adult cancer survivors: results from the National Institutes of Health "All of Us" research program. Cancer Med 2021;10:3646–54.
5. Elsamadicy AA, Wang C, Reeves BC, et al. Socioeconomic and racial/ethnic disparities in perception of health status and literacy in spine oncological patients: insights from the All of Us research program. Spine (Phila Pa 1976) 2023;48:1107–15.
6. Feng J, Symonds EA, Karnes JH. Visualization and quantification of the association between breast cancer and cholesterol in the All of Us research program. Cancer Inform 2023;22:11769351221144132.
7. Hoyt MA, Darabos K, Llave K. Disparities in health-related quality of life among lesbian, gay, and bisexual cancer survivors. J Psychosoc Oncol 2023;41:661–72.
8. Joshi TP, Black TA, Fernandez B, et al. Comorbidities associated with mycosis fungoides: a case-control study in the All of Us database. J Am Acad Dermatol 2023;88:686–8.
9. Keeler Bruce L, Paul P, Kim KK, Kim J, All of Us Research Program Investigators, et al. Family and personal history of cancer in the All of Us research program for precision medicine. PLoS One 2023;18:e0288496.
10. Na J, Zong N, Wang C, et al. Characterizing phenotypic abnormalities associated with high-risk individuals developing lung cancer using electronic health records from the All of Us researcher workbench. J Am Med Inform Assoc 2021;28:2313–24.
11. Ronquillo JG, Lester WT. Precision medicine landscape of genomic testing for patients with cancer in the National Institutes of Health All of Us database using informatics approaches. JCO Clin Cancer Inform 2022;6:e2100152.
12. Ronquillo JG, Lester WT. Pharmacogenomic testing and prescribing patterns for patients with cancer in a large national precision medicine cohort. J Med Genet 2023;60:81–3.
13. Yu J, Sullivan BG, Senthil GN, et al. Prevalence of primary liver cancer is affected by place of birth in Hispanic people residing in the United States: All of Us research program report. Am Surg 2022;88:2565–71.
14. Althoff KN, Schlueter DJ, Anton-Culver H, et al. Antibodies to severe acute respiratory syndrome Coronavirus 2 (SARS-CoV-2) in All of Us research program participants, 2 January to 18 March 2020. Clin Infect Dis 2022;74:584–90.
15. Desine S, Master H, Annis J, et al. Daily step counts before and after the COVID-19 pandemic among

All of Us research participants. JAMA Netw Open 2023;6:e233526.
16. Fan R, Leasure AC, Damsky W, Cohen JM. Association between atopic dermatitis and COVID-19 infection: a case-control study in the All of Us research program. JAAD Int 2022;6:77–81.
17. Gu S, Rajendiran G, Forest K, et al. Drug-induced liver injury with commonly used antibiotics in the All of Us research program. Clin Pharmacol Ther 2023;114:404–12.
18. Lee YH, Liu Z, Fatori D, et al. Association of everyday discrimination with depressive symptoms and suicidal ideation during the COVID-19 pandemic in the All of Us research program. JAMA Psychiatry 2022;79:898–906.
19. Luo W, Baldwin E, Jiang AY, Li S, Yang B, Li H. Effects of housing environments on COVID-19 transmission and mental health revealed by COVID-19 participant experience data from the All of Us research program in the USA: a case-control study. BMJ Open 2022;12:e063714.
20. Miller CN, Althoff KN, Schlueter DJ, All of Us Research Program, et al. Concordance of SARS-CoV-2 antibody results during a period of low prevalence. mSphere. 2022;7:e0025722.
21. Moseley I, Ragi SD, Ouellette S, Rao B. Tinea pedis in underrepresented groups: All of Us database analysis. Mycoses 2023;66:29–34.
22. Pfaff ER, Girvin AT, Crosskey M, N3C and RECOVER Consortia, et al. De-black-boxing health AI: demonstrating reproducible machine learning computable phenotypes using the N3C-RECOVER Long COVID model in the All of Us data repository. J Am Med Inform Assoc 2023; 30:1305–12.
23. Xing LQ, Xu ML, Sun J, et al. Anxiety and depression in frontline health care workers during the outbreak of Covid-19. Int J Soc Psychiatry 2021;67:656–63.
24. Joel MZ, Fan R, Damsky W, Cohen JM. Psoriasis associated with asthma and allergic rhinitis: a US-based cross-sectional study using the All of US research program. Arch Dermatol Res 2023; 315:1823–6.
25. Acosta JN, Leasure AC, Both CP, et al. Cardiovascular health disparities in racial and other underrepresented groups: initial results from the All of Us research program. J Am Heart Assoc 2021;10: e021724.
26. Alonso A, Alam AB, Kamel H, et al. Epidemiology of atrial fibrillation in the All of Us research program. PLoS One 2022;17:e0265498.
27. Chandler PD, Clark CR, Zhou G, All of Us Research Program Investigators, et al. Hypertension prevalence in the All of Us Research Program among groups traditionally underrepresented in medical research. Sci Rep 2021;11:12849.
28. Kurniansyah N, Goodman MO, Khan AT, et al. Evaluating the use of blood pressure polygenic risk scores across race/ethnic background groups. Nat Commun 2023;14:3202.
29. Lee EB, Hu W, Singh K, Wang SY. The association among blood pressure, blood pressure medications, and glaucoma in a nationwide electronic health records database. Ophthalmology 2022;129:276–84.
30. Mensah GA, Jaquish C, Srinivas P, et al. Emerging concepts in precision medicine and cardiovascular diseases in racial and ethnic minority populations. Circ Res 2019;125:7–13.
31. Nagar SD, Pemu P, Qian J, SEEC Consortium, et al. Investigation of hypertension and type 2 diabetes as risk factors for dementia in the All of Us cohort. Sci Rep 2022;12:19797.
32. Renedo D, Acosta JN, Sujijantarat N, et al. Carotid artery disease among broadly defined underrepresented groups: the All of Us research program. Stroke 2022;53:e88–e9.
33. Wang X, Ryu J, Kim J, All of Us Research Program, et al. Common and rare variants associated with cardiometabolic traits across 98,622 whole-genome sequences in the All of Us research program. J Hum Genet 2023;68:565–70.
34. Xue Q, Li X, Wang X, Ma H, Heianza Y, Qi L. Subtypes of type 2 diabetes and incident cardiovascular disease risk: UK Biobank and All of Us cohorts. Mayo Clin Proc 2023;98:1192–204.
35. Almazan E, Schwartz JL, Gudzune KA. Use of medications associated with weight change among participants in the All of Us research programme. Clin Obes 2023;13:e12609.
36. Clark CR, Chandler PD, Zhou G, et al. Geographic variation in obesity at the state level in the All of Us research program. Prev Chronic Dis 2021;18:E104.
37. Giangreco NP, Lina S, Qian J, et al. Pediatric data from the All of Us research program: demonstration of pediatric obesity over time. JAMIA Open 2021;4:ooab112.
38. Abegaz TM, Ahmed M, Sherbeny F, Diaby V, Chi H, Ali AA. Application of machine learning algorithms to predict uncontrolled diabetes using the All of Us research program data. Healthcare (Basel) 2023;11:1138.
39. Chan AX, McDermott JJ, Iv, Lee TC, et al. Associations between healthcare utilization and access and diabetic retinopathy complications using All of Us nationwide survey data. PLoS One 2022;17:e0269231.
40. Delavar A, Radha Saseendrakumar B, Lee TC, et al. Associations between thyroid eye disease and glaucoma among those enrolled in the National Institutes of Health All of Us research program. Ophthalmic Plast Reconstr Surg 2023; 39:336–40.

41. Devineni D, Akbarpour M, Gong Y, Wong ND. Inadequate use of newer treatments and glycemic control by cardiovascular risk and sociodemographic groups in US adults with diabetes in the NIH precision medicine initiative All of Us research program. Cardiovasc Drugs Ther 2022;38:347–57.
42. Fan R, Leasure AC, Maisha FI, Cohen JM, Little AJ. Thyroid disorders associated with lichen sclerosus: a case-control study in the All of Us research program. Br J Dermatol 2022;187:797–9.
43. Lee TC, Radha-Saseendrakumar B, Delavar A, et al. Evaluation of depression and anxiety in a diverse population with thyroid eye disease using the nationwide NIH All of Us database. Ophthalmic Plast Reconstr Surg 2023;39:281–7.
44. Perry AS, Annis JS, Master H, et al. Association of longitudinal activity measures and diabetes risk: an analysis from the national institutes of health All of Us research program. J Clin Endocrinol Metab 2023;108:1101–9.
45. Fan R, Leasure AC, Damsky W, Cohen JM. Migraine among adults with atopic dermatitis: a cross-sectional study in the All of Us research programme. Clin Exp Dermatol 2023;48:24–6.
46. Leasure AC, Acosta JN, Both C, et al. Stroke disparities among nonracial minorities in the All of Us research program. Stroke 2021;52:e488–e90.
47. Renedo D, Acosta JN, Koo AB, et al. Higher hospital frailty risk score is associated with increased risk of stroke: observational and genetic analyses. Stroke 2023;54:1538–47.
48. Ahmed F, Moseley I, Ragi SD, Ouellette S, Rao B. Vitiligo in underrepresented communities: an All of Us database analysis. J Am Acad Dermatol 2023;88:945–8.
49. Ahmed F, Ragi SD, Moseley I, et al. Rosacea diagnosis and prescription patterns in underrepresented groups: an All of Us database analysis. J Am Acad Dermatol 2023;89:384–5.
50. Belzer A, Leasure AC, Cohen JM, Perkins SH. The association of cutaneous squamous cell carcinoma and basal cell carcinoma with solid organ transplantation: a cross-sectional study of the All of Us research program. Int J Dermatol 2023;62:e564–e6.
51. Fan R, Leasure AC, Cohen JM. Association of autoimmune comorbidities with lichen planus: a United States-based case-control study in the All of Us research program. J Am Acad Dermatol 2022;87:1451–3.
52. Fan R, Leasure AC, Damsky W, Cohen JM. Alcohol use disorder among adults with atopic dermatitis: a case-control study in the All of Us research program. J Am Acad Dermatol 2022;87:1378–80.
53. Fan R, Leasure AC, Little AJ, Cohen JM. Lichen sclerosus among women with psoriasis: a cross-sectional study in the All of Us research program. J Am Acad Dermatol 2023;88:1175–7.
54. Hodelin C, Fan R, Damsky W, Cohen JM. The association of atopic dermatitis and chronic rhinosinusitis in adults: a cross-sectional study using the All of Us research program. Int J Dermatol 2023;62:e430–e1.
55. Hong S, Fan R, Cohen JM. Lichen planus is associated with depression and anxiety: a cross-sectional study in the All of Us research program. Arch Dermatol Res 2023;315:1417–9.
56. Joel MZ, Fan R, Cohen JM. Association between psoriasis and celiac disease: a cross-sectional study in the All of Us research program. J Am Acad Dermatol 2023;88:1386–8.
57. Joshi TP, Bancroft A, DeLeon D, et al. Association of atopic dermatitis with substance use disorders: A case-control study in the All of Us research program. J Am Acad Dermatol 2023;89:e237–e8.
58. Joshi TP, Calderara GA, Lipoff JB. Prevalence of pityriasis rosea in the United States: a cross-sectional study using the All of Us database. JAAD Int 2022;8:45–6.
59. Joshi TP, Chen V, Dong JL, et al. Atopic comorbidities associated with granuloma annulare: a case-control study of the All of Us database. J Am Acad Dermatol 2023;89:145–6.
60. Joshi TP, Duruewuru A, Holla S, Naqvi Z, Zhu H, Ren V. Comorbidities associated with lichen planopilaris: a case-control study using the All of Us database. Int J Dermatol 2023;62:e396–e8.
61. Joshi TP, Fernandez B, Friske S, et al. Burden of atopic disease in Black and Hispanic patients with alopecia areata: a case-control study in the All of Us research program. Int J Dermatol 2023;62:e393–e4.
62. Joshi TP, Garcia D, Gedeon F, et al. Epidemiology of alopecia areata in the Hispanic/Latinx community: a cross-sectional analysis of the All of Us database. J Am Acad Dermatol 2023;89:e61–e2.
63. Joshi TP, Zhu H, Naqvi Z, Holla S, Duruewuru A, Ren V. Prevalence of lichen planopilaris in the United States: a cross-sectional study of the All of Us research program. JAAD Int 2022;8:69–70.
64. Leasure AC, Cohen JM. Prevalence of lichen planus in the United States: a cross-sectional study of the All of Us research program. J Am Acad Dermatol 2022;87:686–7.
65. Leasure AC, Cohen JM. Prevalence of eczema among adults in the United States: a cross-sectional study in the All of Us research program. Arch Dermatol Res 2023;315:999–1001.
66. Leasure AC, Damsky W, Cohen JM. Comorbidities associated with granuloma annulare: a case-control study in the All of Us research program. J Am Acad Dermatol 2022;87:197–9.
67. Leasure AC, Damsky W, Cohen JM. Prevalence of granuloma annulare in the United States: a cross-sectional study in the All of Us research program. Int J Dermatol 2022;61:e301–e2.

68. Moseley I, Ragi SD, Lombardi A. Atopic dermatitis in underrepresented groups: an All of Us database analysis. Dermatitis 2022;33:S143–S145.
69. Moseley I, Ragi SD, Ouellette S, Rao B. Condyloma acuminata in under-represented groups: an All of Us database analysis. Sex Transm Infect 2022;98:620–1.
70. Moseley I, Ragi SD, Ouellette S, Rao B. Onychomycosis in underrepresented groups: an All of Us database analysis. Arch Dermatol Res 2023;315:647–51.
71. Moseley IH, George EA, Tran MM, Lee H, Qureshi AA, Cho E. Alopecia areata in underrepresented groups: preliminary analysis of the All of Us research program. Arch Dermatol Res 2023;315:1631–7.
72. Murphy MJ, Heyang M, Fan R, Leasure AC, Damsky W, Cohen JM. Association between uveitis and Lichen planus in the All of Us research program. Arch Dermatol Res 2023;315:2729–30.
73. Murphy MJ, Leasure AC, Damsky W, Cohen JM. Association of sarcoidosis with psoriasis: a cross-sectional study in the All of Us research program. Arch Dermatol Res 2023;315:1439–41.
74. Nock MR, Barbieri JS, Krueger LD, Cohen JM. Racial and ethnic differences in barriers to care among US adults with chronic inflammatory skin diseases: a cross-sectional study of the All of Us research program. J Am Acad Dermatol 2023; 88:568–76.
75. Ragi SD, Lin Z, Moseley I, Ahmed F, Ouellette S, Rao B. Psychiatric comorbidities of Hidradenitis suppurativa in underrepresented groups: a case-control study utilizing the All of Us research program. Arch Dermatol Res 2023;315:1457–9.
76. Tran MM, Moseley IH, George EA, Qureshi AA, Cho E. Examining the burden of psoriasis and psoriatic arthritis in a US adult cohort using the All of Us research program. J Am Acad Dermatol 2023;89:859–62.
77. Rice C, Ayyala DN, Shi H, et al. Sex and racial differences in systemic lupus erythematosus among US adults in the All of Us research program. Arthritis Care Res (Hoboken) 2023;75:2096–106.
78. McDermott JJ, Lee TC, Chan AX, et al. Novel association between opioid use and increased risk of retinal vein occlusion using the National Institutes of Health All of Us research program. Ophthalmol Sci 2022;2.
79. Wu JH, Radha Saseendrakumar B, Moghimi S, et al. Epidemiology and factors associated with cannabis use among patients with glaucoma in the All of Us research program. Heliyon 2023;9:e15811.
80. Venner E, Muzny D, Smith JD, All of Us Research Program Regulatory Working Group, et al. Whole-genome sequencing as an investigational device for return of hereditary disease risk and pharmacogenomic results as part of the All of Us research program. Genome Med 2022;14:34.
81. Acuff K, Delavar A, Radha Saseendrakumar B, Wu JH, Weinreb RN, Baxter SL. Associations between socioeconomic factors and visit adherence among patients with glaucoma in the All of Us research program. Ophthalmol Glaucoma 2023; 6:405–12.
82. Baxter SL, Saseendrakumar BR, Paul P, All of Us Research Program Investigators, et al. Predictive analytics for glaucoma using data from the All of Us research program. Am J Ophthalmol 2021;227:74–86.
83. Chan AX, Radha Saseendrakumar B, Ozzello DJ, et al. Social determinants associated with loss of an eye in the United States using the All of Us nationwide database. Orbit 2022;41:739–44.
84. Delavar A, Radha Saseendrakumar B, Weinreb RN, Baxter SL. Racial and ethnic disparities in cost-related barriers to medication adherence among patients with glaucoma enrolled in the National Institutes of Health All of Us research program. JAMA Ophthalmol 2022;140:354–61.
85. Delavar A, Saseendrakumar BR, Weinreb RN, Baxter SL. Healthcare access and utilization among glaucoma patients in a nationwide cohort. J Glaucoma 2023;32:40–7.
86. Sekimitsu S, Shweikh Y, Zebardast N. Effect of visual impairment on depression and anxiety during the COVID-19 pandemic in the United States. Can J Ophthalmol 2022.
87. Almazan E, Yenokyan G, Ng K. Systemic diseases associated with a diagnosis of achalasia: a case-control study with the All of Us research program. Eur J Intern Med 2022;104:125–7.
88. Weidner A, Asif I. Shaping the future of family medicine research: the 2023 National Family Medicine Research Summit. Ann Fam Med 2024;22:72–4.
89. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. 2013. Available at: http://www.R-project.org/.
90. Van Rossum G, Drake FL, Jr. The python language reference. Python Software Foundation: Wilmington, DE, USA. 2014.
91. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in Neural Inf Process Syst 2017;30.
92. Dettmers T, Pagnoni A, Holtzman A, Zettlemoyer L. Qlora: efficient finetuning of quantized llms. arXiv preprint arXiv 2023; 230514314.
93. Frantar E, Ashkboos S, Hoefler T, Alistarh D. Gptq: accurate post-training quantization for generative pre-trained transformers. arXiv preprint arXiv 2022; 221017323.
94. Gerganov G. GitHub: ggerganov/ggml 2023 [cited 2023-09-28]. Available at: https://github.com/ggerganov/ggml.
95. Department of Health and Human Services. Code of Federal Regulations Title 45 Part 46 Protection

of Human Subjects. Available at: https://www.ecfr.gov/current/title-45/part-46.

96. Barr PB, Bigdeli TB, Meyers JL. Prevalence, comorbidity, and sociodemographic correlates of psychiatric disorders reported in the All of Us research program. JAMA Psychiatry 2022;79:622–8.

97. Accreditation Council for Graduate Medical Education. ACGME program requirements for graduate medical education in family medicine 2023 [cited 2023 2023-11-14]. Available at: https://www.acgme.org/globalassets/pfassets/programrequirements/120_familymedicine_2023.pdf.