DOCUMENT SUMMARY

This special report from the New England Journal of Medicine describes the "All of Us" Research Program, a massive, federally-funded initiative to build a health research cohort of at least one million people in the United States. It is critically relevant to Enlitens because it explicitly states that past medical research has been hindered by a lack of diversity, leading to decreased generalizability and incorrect interpretations. The program's core mission to recruit a cohort that reflects the diversity of the U.S.—with over 80% of its initial participants from groups historically underrepresented in research—provides powerful, large-scale validation for Enlitens' argument against using narrow, non-representative data (like that from standardized tests) to make judgments about individuals.

FILENAME

All_of_Us_Investigators_2019_Large_Scale_Diverse_Cohort_evidence_for_diversity_in_research

METADATA

- **Primary Category**: RESEARCH
- **Document Type**: research_article
- **Relevance**: Supporting
- **Key Topics**: research_diversity, health_disparities, large_cohort_study, social_determinants_of_health, participant_engagement, EHR, data_bias
- **Tags**: #AllOfUs #ResearchDiversity #HealthEquity #BigData #UnderrepresentedPopulations #NIH #PrecisionMedicine #BiasInResearch

CRITICAL QUOTES FOR ENLITENS

- "Many of these cohorts, however, are small, lack diversity, or do not provide comprehensive phenotype data."
- "More than 80% of these participants are from groups that have been historically underrepresented in biomedical research."
- "However, many efforts have been hampered by an inadequate sample size and a lack of diversity among participants,[1] restrictive policies regarding data access, or failure to capture genotype and phenotype data comprehensively."
- "Collectively, these challenges have slowed the pace of medical discovery, decreased the generalizability of research findings, hindered reproducibility, and led to incorrect interpretations."
- "Population-based research, which requires large sample sizes and highly granular phenotypic data, benefits from access to populations of patients from various ancestries."

- "The All of Us Research Program seeks to recruit persons in demographic categories that have been and continue to be underrepresented in biomedical research; such persons typically have relatively poor access to good health care."
- "Race, ethnic group, age, sex, gender identity, sexual orientation, disability status, access to care, income, educational attainment, and geographic location are therefore taken into account."
- "Because racial and ethnic identities are more than a genetic construct, we seek to capture other social and behavioral determinants of health."
- "EHR data can be fragmented, incomplete, and inaccurate 32,36; however, in certain contexts, such data have already proved to be a powerful and efficient discovery tool."

KEY STATISTICS & EVIDENCE

- **Program Goal**: To enroll a diverse group of at least 1 million persons in the United States.
- **Recruitment Status (as of July 2019)**: More than 175,000 participants had contributed biospecimens , and more than 230,000 total participants were enrolled.
- **Diversity Statistics (as of July 2019)**:
    - More than 80% of the 175,000 participants who contributed biospecimens are from groups historically underrepresented in biomedical research.
    - Among core participants, 51% are nonwhite.
    - Among core participants, 80% meet the program's definition of being underrepresented in biomedical research.
- **Diversity Recruitment Targets**:
    - The target percentage of persons in racial and ethnic minorities is more than 45%.
    - The target percentage of persons in underrepresented populations is more than 75%.
- **EHR Data Collection (as of July 2019)**: EHR data on more than 112,000 participants from 34 sites have been collected.
- **Recruitment Rate**: Approximately 3100 core participants per week.
- **Projected Goal Completion**: Expects to have enrolled 1 million core participants by approximately 2024.
- **Projected Genetic Findings**:
    - An estimated 30,000 persons will receive actionable findings for ACMG genes.
    - More than 90% of participants may learn of actionable pharmacogenomic variants.
- **Funding**: As of 2015, Congress has allocated $1.02 billion to the program, with the 21st Century Cures Act authorizing an additional $1.14 billion through 2026.

METHODOLOGY DESCRIPTIONS

The All of Us Research Program utilizes a comprehensive protocol designed for large-scale, longitudinal data collection from a diverse population.

- **Recruitment and Consent**:

    - Participants must be 18 years or older with the capacity to provide informed consent. Protocols are in development to include children, adolescents, and cognitively impaired persons.
    - Enrollment occurs digitally through a website or smartphone app.
    - The consent process includes explanatory videos with brief text, iconography, and formative questions to ensure understanding. All content is targeted to a fifth-grade reading level and is available in English and Spanish.
    - Recruitment happens through a network of over 340 sites, including health care provider organizations (regional medical centers, federally qualified health centers, Veterans Health Administration) and a "direct-volunteer" route for those unaffiliated with these centers.
- **Data Collection Elements**: The program gathers multiple streams of data.
    - **Health Surveys**: Initial surveys cover sociodemographics, overall health, lifestyle, and substance use. Subsequent modules cover personal and family medical history and access to health care. All survey questions were evaluated with cognitive and online testing in diverse populations before use.
    - **Electronic Health Records (EHRs)**: With participant authorization, the program captures structured data including billing codes, medication history, laboratory results, vital signs, and encounter records. This will be expanded to include narrative documents. Data is structured according to a common data model.
    - **Physical Measurements**: Per-protocol measurements include blood pressure, heart rate, weight, height, body-mass index, and hip and waist circumferences.
    - **Biospecimens**: Blood and urine samples are collected for DNA, RNA, cell-free DNA, serum, and plasma. If blood cannot be obtained, saliva specimens are used.
    - **Digital Health Technology**: Participants can share data from compatible devices like Fitbit. The program plans to expand support for other devices.
    - **Future Data Sources**: The program plans to link to other datasets like national death indexes, pharmacy data, health care claims data, and geospatially linked environmental data (weather, air quality, etc.).
- **Data Access for Researchers**:
    - To promote broad access and avoid long delays, the program uses a cloud-based environment where approved researchers can use web-based tools to analyze data without downloading it.
    - The program uses a "passport model," where researchers are approved for data access to study any topic that meets the criterion for allowable use, rather than requiring project-specific applications.
    - All identifying information is removed from participant data available to researchers.

POPULATION-SPECIFIC FINDINGS

This document describes the *methodology* for recruiting diverse populations rather than presenting research findings *about* them. The core of the program's design is to intentionally include groups that have been historically underrepresented, making the methodology itself a key finding for Enlitens.

- **Focus on Underrepresented Groups**: "The All of Us Research Program seeks to recruit persons in demographic categories that have been and continue to be

underrepresented in biomedical research; such persons typically have relatively poor access to good health care."

- **Broad Definition of Diversity**: "Race, ethnic group, age, sex, gender identity, sexual orientation, disability status, access to care, income, educational attainment, and geographic location are therefore taken into account."
- **Prioritization**: "Persons in underrepresented populations who are enrolled in the program will be prioritized for physical measurements and biospecimen collections."
- **Beyond Genetics**: "Because racial and ethnic identities are more than a genetic construct, we seek to capture other social and behavioral determinants of health."
- **Community Engagement**: The program has funded a network of 22 community partners to engage diverse populations and providers, guided by pilot studies.

PRACTICAL APPLICATIONS

The practical application for Enlitens is the program's model for comprehensive, multi-modal data collection, which stands in stark contrast to the single data point of a standardized test. The types of data collected provide a blueprint for what a holistic assessment could entail.

Table 1. Data Available to Researchers from the All of Us Cohort.

| Data Source | Details | | :--- | :--- | | **Current sources** | | | Health surveys | Initial surveys include information on sociodemographic characteristics, overall health, lifestyle, and substance use, with subsequent modules covering personal and family medical history and access to health care. |

| Physical measurements | Per-protocol measurements include blood pressure, heart rate, weight, height, body-mass index, and hip and waist circumferences. |

| Biospecimens & Electronic health records | Blood and urine samples are tested for DNA, RNA, cell-free DNA, serum, and plasma. If blood specimens cannot be obtained, saliva specimens are obtained. Initial capture of structured data includes billing codes, medication history, laboratory results, vital signs, and encounter records from health care provider organizations. Records will be expanded to include narrative documents. Pilot studies are testing data collection through Sync for Science and other health data aggregators. |

| Digital health information | Data can be captured from compatible participant-owned devices such as Fitbit. Pilot studies of other devices and linkage to health apps are being explored. |

| **Future sources** | | | Health surveys | Additional modules, including surveys regarding social behavioral determinants of health, are under development. |

| Bioassays | Pilot studies for genotyping and whole-genome sequencing are expected to begin by early 2020. Additional pilot studies of bioassays are planned. |

| Health care claims data. | Systems for the use of claims data, including billing codes and medication data, are under development. |

| Geospatial and environmental data | These data include geospatial linkage to measures such as weather, air quality, pollutant levels, and census data. Assays and sensor-based measurements of exposure are under consideration. |

| Other sources | Voluntary contributions of data from social networks (e.g., Twitter feeds) and additional biospecimen collections are under consideration. |