# Understanding Precision Medicine through the NIH *All of Us* Research Program and NCI Cancer Research Data Commons

**Jay G. Ronquillo**[1,2], **William T. Lester**[3,4]
[1]Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD
[2]Office of Data Science Strategy, National Institutes of Health, Bethesda, MD
[3]Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA
[4]Harvard Medical School, Boston, MA

## Background

The National Institutes of Health (NIH) *All of Us* (AoU) Research Program is a nationwide initiative focused on collecting vast biomedical and real-world data from 1 million+ patients in order to accelerate the research and practice of precision medicine.(1) The AoU dataset is harmonized to the Observational Health Data Sciences and Informatics (OHDSI) Observational Medical Outcomes Partnership (OMOP) Common Data Model. Similarly, the National Cancer Institute (NCI) Cancer Research Data Commons (CRDC) is a cloud-based data ecosystem to catalyze large-scale cancer research.(2) This study investigates the current state of precision medicine for cancer patients by leveraging both the NIH AoU program and NCI CRDC.

## Methods

Using a recent version of the AoU dataset (December 2020), we defined genomic testing and cancer categories, respectively, using the Logical Observation Identifiers Names and Codes (LOINC) database and the NCI Surveillance, Epidemiology and End Results (SEER) Site Modules. An AoU participant was defined as having cancer if they reported at least one cancer diagnosis in their medical history survey. Using cloud-based Jupyter Notebooks (Python 3.7, R 4.0.3), all cancer-related definitions and mappings were created in the NCI Cancer Research Data Commons, while data extraction, integration, and analysis was performed in the AoU Researcher Workbench (Figure 1).

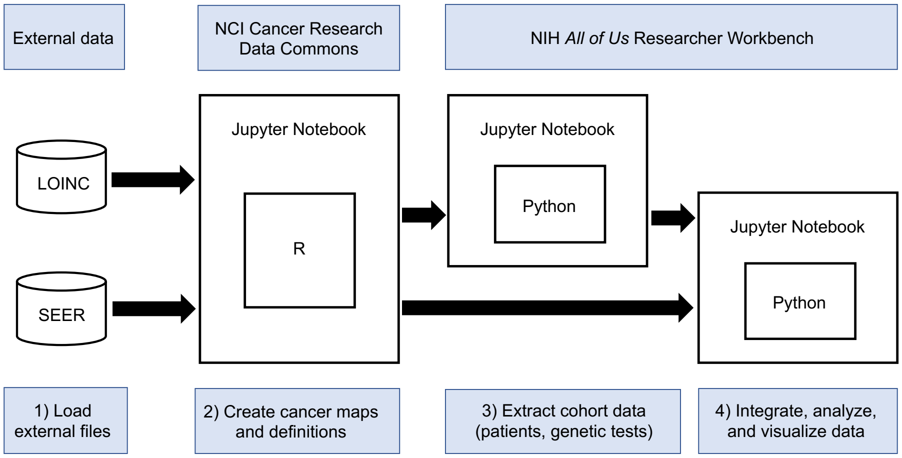
External data
NCI Cancer Research Data Commons
NIH *All of Us* Researcher Workbench
LOINC
SEER
Jupyter Notebook
R
Jupyter Notebook
Python
Jupyter Notebook
Python
1) Load external files
2) Create cancer maps and definitions
3) Extract cohort data (patients, genetic tests)
4) Integrate, analyze, and visualize data

Figure 1. Biomedical informatics pipeline used to extract data for *All of Us* cancer patients.

## Results

There were a total of 20333 AoU participants who reported having at least one cancer diagnosis (Figure 2). By age group, this included 136 (0.7%) patients between 18-29 years, 615 (3.0%) between 30-39 years, 1239 (6.1%) between 40-49 years, 2792 (13.7%) between 50-59 years, 6193 (30.5%) between 60-69 years, 7428 (36.5%) between 70-79 years, and 1930 (9.5%) between 80-89 years. The cancer cohort was composed of 12282 (60.4%) females, 7864 (38.7%) males, and 187 (0.9%) not specified/other. By race, there were 213 (1.0%) Asian, 697 (3.4%) Black or African American, 18201 (89.5%) White, and 1222 (6.0%) not specified/other. By ethnicity, there were 732 (3.6%) Hispanic or Latino, 19167 (94.3%) Not Hispanic or Latino, and 434 (2.1%) not specified/other.

There were 622 (3.2%) cancer patients who received some form of genomic testing, with the most common tests focused on specific gene mutations (totaling 945 tests and affecting 416 patients). The most frequently assessed genes included the following: HBB, F5, F2, JAK2, MTHFR, CFTR, and HFE.

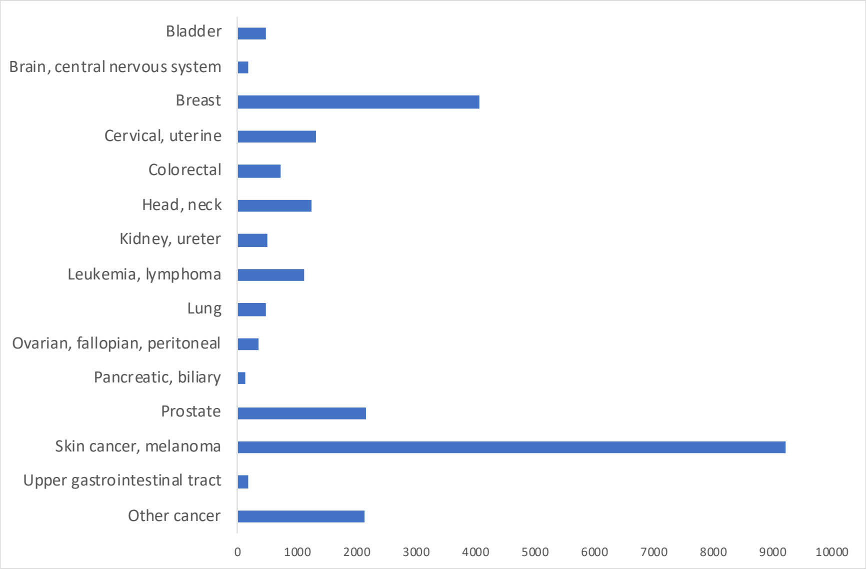

Figure 2. Distribution of cancer diagnoses reported by *All of Us* participants.

## Conclusion

Biomedical informatics pipelines can be developed to extract, integrate and analyze cancer-related data from diverse cloud-based platforms for precision medicine research.

## References/Citations

1. Ramirez AH, Gebo KA, Harris PA. Progress With the All of Us Research Program Opening Access for Researchers. JAMA. 2021;June 11.
2. Hinkson I V., Davidsen TM, Klemm JD, Kerlavage AR, Kibbe WA. A comprehensive infrastructure for big data in cancer research: Accelerating cancer research and precision medicine. Front Cell Dev Biol. 2017;5(83).