

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

# Jav G. Ronguillo, MD, MPH, MMSc, MEng<sup>1,2</sup> and William T. Lester, MD, MS<sup>3,4</sup>

<sup>1</sup>Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD

<sup>2</sup>Office of Data Science Strategy, National Institutes of Health, Bethesda, MD

<sup>3</sup>Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA

<sup>4</sup>Harvard Medical School, Boston, MA

Email: iav.ronguillo@nih.gov

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

## Background

- · The National Institutes of Health (NIH) All of Us (AoU) Research Program is collecting vast electronic health record (EHR) data from 1 million+ patients across the country to catalyze the practice of precision medicine.
- The National Cancer Institute (NCI) Cancer Research Data Commons (CRDC) is a cloud-based data ecosystem containing both cancer-related data and analytics tools to catalyze large-scale precision oncology.2
- Focusing on cancer patients, this study investigates the current state of precision medicine and precision oncology by leveraging both the NIH AoU program and NCI CRDC.

### Methods

- Data source: most recent (December 2020) release of the AoU dataset, made available to researchers as Registered Tier data following the OHDSI OMOP Common Data Model
- Cancer cohort: any patient reporting at least 1 cancer diagnosis in their Medical History survey, and grouped following NCI Surveillance, Epidemiology and End Results (SEER) site classifications

Results

Specified/Other (2.1%)

Most common genes:

· Most common testing categories:

Genomic testing

- Genomic testing: defined by list of relevant LOINC codes focused on assessing human genetic or genomic information
- Cloud-based Jupyter Notebook (R 4.0.3) in the NCI Cancer Research Data Commons created all cancer definitions and mapping files
- Cloud-based Jupyter Notebook (python 3.7) in AoU Researcher Workbench used for data extraction, integration, and analysis
- Full biomedical informatics pipeline described in Figure 1



#### Summary and Conclusions

- Biomedical informatics pipelines can extract, integrate and analyze cancer-related data from diverse cloudbased platforms for precision medicine research.
- More researchers are needed to leverage federal cloudbased precision medicine resources in order to realize the full potential of precision medicine.

For more information or to request access to data and/or cloud-based tools:

- NCI Cancer Research Data Commons
  - https://datascience.cancer.gov/
- NIH All of Us Research Program
  - https://www.researchallofus.org/

#### References

- 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), doi:10.3389/fcell.2017.00083

### Acknowledgments

This study was supported by general cloud credits via the National Institutes of Health (NIH) Science and Technology Research Infrastructure for Discovery, Experimentation, and Sustainability (STRIDES) Initiative and the National Cancer Institute (NCI) Cancer Research Data Commons Cloud Resources

The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #; AOD 16037: Federally Qualified Health Centers: HHSN 263201600085U: Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163: Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277: 3 OT2 OD025315: 1 OT2 OD025337: 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants

The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.