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

Jay G. Ronquillo, MD, MPH, MMSc, MEng1,2 and William T. Lester, MD, MS3,4
1Center for Biomedical Informatics and Information Technology, National Cancer Institute, Rockville, MD
2Office of Data Science Strategy, National Institutes of Health, Bethesda, MD
3Laboratory of Computer Science, Massachusetts General Hospital, Boston, MA
4Harvard Medical School, Boston, MA
Email: jay.ronquillo@nih.gov

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

Results

• 20,333 cancer patients with diverse cancers (Figure 2)
  • Female (60.4%), Male (38.7%), Not Specified/Other (0.9%)
  • Asian (1.0%), Black or African American (3.4%), White (89.5%), Not Specified/Other (0.0%)
  • Hispanic or Latino (3.6%), Not Hispanic or Latino (94.3%), Not Specified/Other (2.1%)
• Genomic testing
  • 622 (3.2%) of cancer patients received some form of genomic testing
  • Most common testing categories:
    • Gene mutations (945 tests, 416 patients)
    • Molecular pathology (403 tests, 164 patients)
    • Gene translocation (288 tests, 42 patients)
  • Most common genes:
    • HBB, F5, F2, JAK2, MTHFR, CFTR, HFE

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

Figure 2. Distribution of AoU participant-reported cancer diagnoses.

Summary and Conclusions

• Biomedical informatics pipelines can extract, integrate and analyze cancer-related data from diverse cloud-based platforms for precision medicine research.
• More researchers are needed to leverage federal cloud-based 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.research.aou.org

References


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 CT2 OD020844; 1 CT2 OD020854; 1 CT2 OD020857; 1 CT2 OD020856; 1 OT2 OD020855; 1 OT2 OD020854; 1 OT2 OD020851; 1 OT2 OD020855; 1 CT2 OD0210853; 1 CT2 OD0210857; Federally Qualified Health Centers: N01-CH-13016; National Cancer Institute: N01-CH-13017; NCI Cancer Research Data Commons Cloud Resources.

This study was supported by general cloud credits via the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 CT2 OD020844; 1 CT2 OD020854; 1 CT2 OD020857; 1 CT2 OD020856; 1 OT2 OD020855; 1 OT2 OD020854; 1 OT2 OD020851; 1 OT2 OD020855; 1 CT2 OD0210853; 1 CT2 OD0210857; Federally Qualified Health Centers: N01-CH-13016; National Cancer Institute: N01-CH-13017; NCI Cancer Research Data Commons Cloud Resources.

This study was supported by general cloud credits via the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 CT2 OD020844; 1 CT2 OD020854; 1 CT2 OD020857; 1 CT2 OD020856; 1 OT2 OD020855; 1 OT2 OD020854; 1 OT2 OD020851; 1 OT2 OD020855; 1 CT2 OD0210853; 1 CT2 OD0210857; Federally Qualified Health Centers: N01-CH-13016; National Cancer Institute: N01-CH-13017; NCI Cancer Research Data Commons Cloud Resources.

This study was supported by general cloud credits via the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 CT2 OD020844; 1 CT2 OD020854; 1 CT2 OD020857; 1 CT2 OD020856; 1 OT2 OD020855; 1 OT2 OD020854; 1 OT2 OD020851; 1 OT2 OD020855; 1 CT2 OD0210853; 1 CT2 OD0210857; Federally Qualified Health Centers: N01-CH-13016; National Cancer Institute: N01-CH-13017; NCI Cancer Research Data Commons Cloud Resources.