Learning important common data elements from shared study data: The All of Us program analysis

Craig Mayer and Vojtech Huser

Lister Hill National Center for Biomedical Communication, National Library of Medicine, NIH Bethesda, MD

National Library of Medicine

Introduction

- All of Us
  - Large scale research program
  - Platform for numerous observational research analyses
  - Combines survey questionnaires, research visit, and Electronic Health Record
  - Uses the OMOP CDM
  - Formalized developed data elements in PPI terminology
  - 329,070 at the time of analysis
  - Three tiers: Public, Registered, and Controlled

Objective

- Identify Common Data Elements and Unique Data Elements in the research data elements
  - Including both Case Report forms and research visit
  - Original source of CDEs
  - How many CDEs crossover between research and EHR

Methods

- Analyzed All of Us data in the registered tier from the third quarter of 2021

- Reviewed Data Elements and Permissible Values

- Defined Common Data Elements and Common Permissible values as concepts originating from outside the All of Us PPI terminology

- Used source (PPI/PM) to identify research data elements or HER

- Identified crossover between research and EHR

Data Elements

- 1,033 Total Data Elements
- 164 CDEs (15.9%)

  - 103 LOINC
  - 60 SNOMED CT

Permissible Values

- 932 Total Permissible Values
- 4,592 DE-PV combinations
- 575 (61.7%) PVs from standard terminologies

Vocabulary ID	Distinct Values	Percentage of Values	Elements with value terminology	Percent of Elements with values terminology
LOINC	194	20.82%	748	72.41%
PPI	344	36.91%	768	74.35%
SNOMED CT	357	38.30%	338	32.72%
AoU_General	13	1.39%	12	1.16%
None	1	0.11%	68	6.58%
ICD10CM	9	0.97%	1	0.10%
ICD9CM	11	1.18%	1	0.10%
UCUM	3	0.32%	2	0.19%

Data Elements

Concept ID	Concept Name	Vocabulary ID	Concept Code	Event Count	CRF*
3027018	Heart rate	LOINC	8867-4	13,609,305	
3025315	Body weight	LOINC	29463-7	3,215,985	
40766240	Are you covered by health insurance or some other kind of health care plan [PhenX]	LOINC	63513-6	931,119	The Basics
1333104	In the past month, to cope with social distancing and isolation, are you doing any of the following? Select all that apply.	PPI		918,790	COPE
4214956	History of clinical finding in subject	SNOMED	417662000	597,652	
1585636	Recreational Drug Use: Which Drugs Used	PPI		541,557	Lifestyle
40771103	How often do you have a drink containing alcohol	LOINC	68518-0	402,030	Lifestyle
40771090	Current occupational status [SAMHSA]	LOINC	68505-7	344,088	The Basics
1585389	Health Insurance: Health Insurance Type	PPI		339,977	
1586140	Race: What Race Ethnicity	PPI		336,278	The Basics
1585370	Home Own: Current Home Own	PPI		329,038	The Basics
1585766	Overall Health: Medical Form Confidence	PPI		329,038	Overall Health

Initiative CDEs

- 87 initiative CDEs

- 15 data collection initiatives

Initiative	Element Count	Percent of Initiative CDEs
PhenX	17	19.54%
PROMIS	15	17.24%
Reported.PHQ	14	16.09%
Perceived Stress Scale-10	10	11.49%
MOS Social Support Survey	9	10.34%
HHS.ACA Section 4302	6	6.90%
SAMHSA	5	5.75%
IPAQ	3	3.45%
UCLA Loneliness Scale v3	2	2.30%

EHR and Research Crossover

- 64 DEs crossover between EHR and Research

Element Concept ID	Element	Element Vocabulary ID	Element Concept Code	Event Count	Percent of Total Participants	Percent of Events from CRFs
3027018	Heart rate	LOINC	8867-4	13 609 305	82.09%	5.90%
3012888	Diastolic blood pressure	LOINC	8462-4	9 758 808	81.85%	8.15%
3038553	Body mass index (BMI) [Ratio]	LOINC	39156-5	2 421 692	80.52%	10.89%
4214956	History of clinical finding in subject	SNOMED	417662000	597 652	45.63%	29.81%
40771103	How often do you have a drink containing alcohol	LOINC	68518-0	402 030	86.21%	99.06%
40771104	How many standard drinks containing alcohol do you have on a typical day [SAMHSA]	LOINC	68519-8	317 050	69.87%	98.83%
40766929	How many cigarettes do you smoke per day now [PhenX]	LOINC	64218-1	127 781	37.76%	97.18%
4179963	Family history of breast cancer	SNOMED	429740004	51 401	3.91%	5.32%

Conclusions

- Limited number of CDEs: 15.9%
  - Possibilities for increased usage moving forward
  - Opportunity for semantic mapping of UDEs

- Select EHR and Research crossover: 64 Data Elements
  - Allows for routine monitoring and follow up

Links & References

- Publication:
  - Mayer CS, Huser V (2023) Learning important common data elements from shared study data: The All of Us program analysis. PLOS ONE 18(7): e0283601. https://doi.org/10.1371/journal.pone.0283601

- Results Repository
  - https://github.com/lhncbc/CRI/tree/master/AoU/CDE
  - Updated Results
  - Controlled Tier Results

- R4aou: R package for All of Us analysis
  - CRI/AoU/r4aou at master · lhncbc/CRI · GitHub

Contact: craig.mayer2@nih.gov

NIH National Library of Medicine