Quantifying Condition Completeness using Medications in the *All of Us* Research Program

Lina Sulieman, PhD[1], Xinzhuo Jiang, MS[2], Joshua Smith, PhD[1], Karthik Natarajan, PhD[2], Paul Harris, PhD[1]

[1]Vanderbilt University Medical Center, Nashville, TN; [2]Columbia University Irving Medical Center, New York, NY

Background

The secondary usage of Electronic Health Records (EHRs) has the potential to advance discovery and innovation. However, the credibility of research can be influenced by the data completeness[1]. EHR Data completeness encompasses multiple dimensions, such as breadth and density. Existing completeness metrics typically focus on assessing the presence of expected clinical events within patient records[2]. Such general metrics often lack specificity in quantifying completeness for specific domains, such as medication records and their associated indications.

Medications are prescribed to treat specific conditions, and various knowledgebases exist to map medications to their conditions. For example, the National Drug File-Reference Terminology (NDF-RT) provides detailed information on drugs, their ingredients, and indicated conditions[3]. Smith developed the Drug Evidence Base (DEB), a medication-indication knowledgebase combining data from MEDLINE, MedlinePLUS, NDF-RT, SIDER, and DrugBank[4]. Gaps between medication records and corresponding conditions can result from mapping errors, EHR fragmentation, or missing non-billable codes. Such gaps may signal incomplete condition documentation, potentially affecting analyses in large biomedical datasets like the All of Us Research Program.

The *All of Us* Research Program is a national initiative aimed at building a diverse biomedical research repository. The program aggregates EHR data from 64 healthcare provider organizations (HPOs) in OMOP common data model format. This study aims to quantify the missingness of condition documentation in EHRs within the All of Us Research Program, leveraging medication knowledgebases to assess gaps and inconsistencies.

Methods

We utilized OMOP concept relationship table to extract NDF-RT relationships for medications and conditions. The *All of Us* Research Program standardized medications to RxNorm and conditions to SNOMED. We mapped NDF-RT medications to RxNorm codes and NDF-RT conditions to SNOMED, ICD-9, and ICD-10 codes. These condition codes were subsequently mapped to phecodes to facilitate analysis. Using the OMOP ancestor and relationship tables, we extracted RXNorm ingredients listed in DEB2 and medications associated with those ingredients. We mapped DEB2 conditions to SNOMED, ICD-9, and ICD-10. Medication-condition pairs were defined as conditions treated by medications documented in either OMOP or DEB2.

We analyzed the All of Us curated dataset version-8 (CDR-8) to extract medication-condition pairs. To assess condition documentation completeness, participants needed at least two entries for a medication and one for a condition it treats (e.g., two Albuterol for any asthma or bronchospasms entries). Completeness was defined as the proportion of participants with both medication and corresponding condition entries, relative to those with the medication alone. To examine if indication strength affects missingness, we varied the evidence threshold for medication-condition links and measured its impact on completeness. Conditions were also mapped to Phecode groups to visualize completeness across categories.

Results

DEB2 included 6,431 medication-conditions relationships for 1130 medications and 1,222 conditions. We used relationship and ancestor tables since we noticed that the ancestor table does not include all medications and ingredients listed in the relationship table (See Table 1). For example, Menthol ingredient had only 16 descendants in ancestor while using “RXnorm is ing of” in the relationship table provided 1483 medications.

In *All of Us* CDR-8, 362,134 participants had medications or conditions where 326,073 participants (90.04%) had both drugs and conditions, and 8,089 participants had medications without any condition documentation. Using medication-condition from one to six resources, we calculated condition completeness using 1,501, 971, 699, 412, and 154 pairs (Table 2). Condition completeness of 0.9 or higher was associated with the lowest number of medication-condition pairs, with 220 pairs supported by evidence from one resource. When using one resource, 868 medication-condition pairs had condition completeness of 0.5 or higher. In contrast, stronger evidence from six resources resulted in 137 medication-condition pairs with condition completeness proportions of 0.5 or higher. For completeness proportions of 0.9 or higher supported by six resources, only 32 medication-condition pairs were identified. Table 2 illustrates the relationship between the strength of evidence and condition completeness, highlighting that higher evidence thresholds correspond to fewer medication-condition pairs with high completeness proportions. Using four resources had higher median of condition completeness compared to using two resources as Figure 2 depicts. Using four resources, the condition completeness of medication that were prescribed for more than 5000 participants ranged between 0.2 to 1 while condition completeness for medications prescribed for less than 125 participants had a wider completeness range between 0 to 1 (Figure 2). Common conditions treated by common medications such as ibuprofen, acetaminophen and albuterol had condition completeness that were higher than 0.8 as Table 3 shows. Propofol and midazolam that are used to relieve anxiety before a procedure and documented as treatment for epilepsy, dementia or anxiety had 0.232 and 0.373 completeness proportion respectively.

The completeness of conditions varied based on Pheocde category as Figure 2 shows. Cardiovascular category had the highest mean of completeness with a value of 0.811 followed by symptoms category with mean of completeness of 0.80. Genetics and pregnancy had the lowest mean of condition completeness where the mean completeness values were 0.61 and 0.64 respectively. Figure 2 shows the condition completeness grouped by phecode category. Focusing on specific chronic diseases, heart diseases and diabetes diseases had a completeness median higher than 0.8 as boxplot in Figure 3 depicts. Cancer diseases had the lowest median of documentation completeness with 0.5 value.

Conclusion

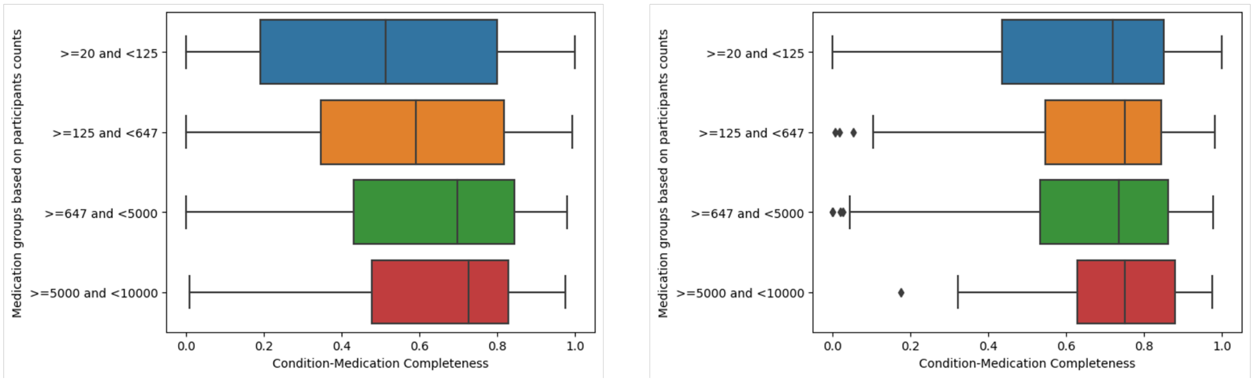
Our pilot study proposes a method to quantify conditions documentation completeness using medication-condition evidence reported in OMOP and DEB2. This is the first attempt, to our knowledge, to quantify conditions missingness at scale. Genetics and pregnancy codes had a high missingness which might happen due to mapping errors or participants receiving care at other specialized healthcare organizations while chronic diseases such as cardiovascular diseases and asthma had high documentation completion rates. Our generalizable method can be applied to any OMOP dataset, such as All of Us Research Program, to identify issues that lead to missingness in participants' records and assess condition completeness. Moreover, this method covers a wide range of conditions and medications which can provide completeness and plausibility metrics, one of the challenging data quality dimensions.

Table 1. Differences between using union of ancestor and relationship, ancestor only and relationship only tables to extract medications with ingredients

Drug ingredient concept id	**Drug ingredient name**	**Union ancestor and relationship count**	**Ancestor only count**	**Relationship only count**
901656	Menthol	1497	16	1483
19009540	Vitamin E	1416	19	1398
967823	Sodium chloride	1113	51	1063
964407	salicylic acid	1076	41	1036
917006	Benzocaine	736	20	717
908523	mineral oil	205	5	201
979096	zinc acetate	141	3	139
967861	magnesium citrate	137	1	137
19101604	cod liver oil	128	2	127
1237049	theophylline	120	120	0

Table 2. The number of condition-medication pairs stratified by number of resources documenting that relationship and condition completeness ratio in *All of Us* Research Program

Condition completeness ratio	**One Resource 1501 medications**	**Two Resources 1501 medications**	**Three Resources 971 medications**	**Four Resources 699 medications**	**Five Resources 412 medications**	**Six Resources 154 medications**
>=0.5	868 (57.83)	868 (57.83)	649 (66.84)	528 (75.54)	342 (83.01)	137 (88.96)
>=0.6	738 (49.17)	738 (49.17)	559 (57.57)	466 (66.67)	306 (74.27)	122 (79.22)
>=0.7	611 (40.71)	611 (40.71)	463 (47.68)	397 (56.8)	263 (63.83)	105 (68.18)
>=0.8	423 (28.18)	423 (28.18)	307 (31.62)	257 (36.77)	173 (41.99)	70 (45.45)
>=0.9	220 (14.66)	220 (14.66)	148 (15.24)	125 (17.88)	81 (19.66)	32 (20.78)



Using two resources **Using Four Resources**

Figure 1. Box plot for medication-condition completeness grouped by number of participants

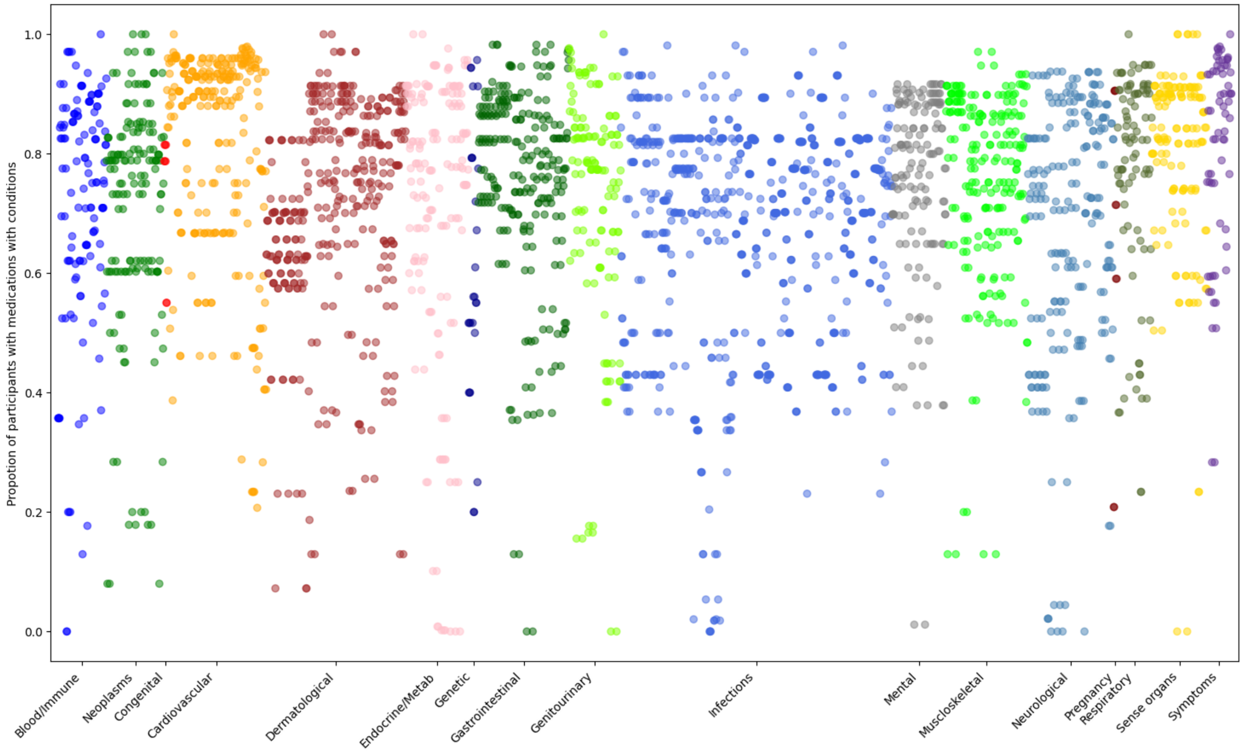


Figure 2. Medication-Condition completeness grouped by conditions Phecode – at least four resources

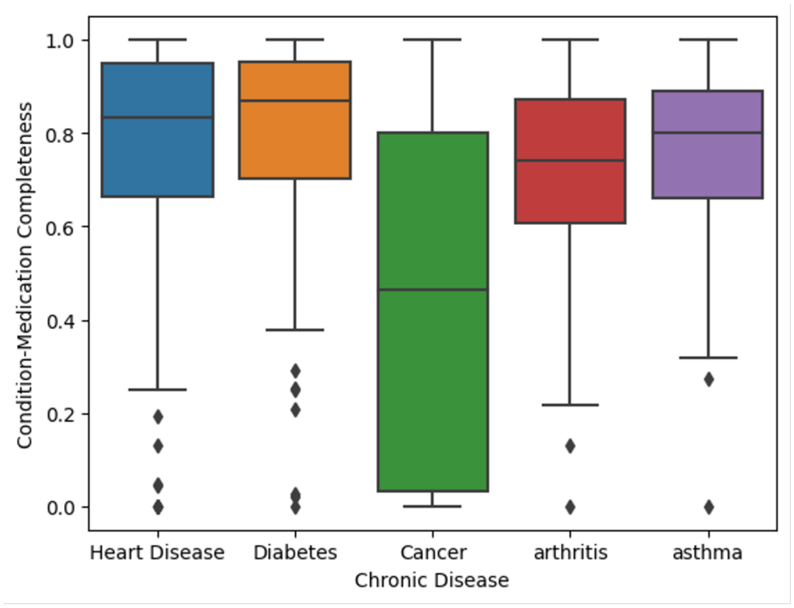


Figure 3. Boxplot for completeness of common chronic condition including heart disease, diabetes, cancer, arthritis and asthma

Table 3. The condition completeness for medications with highest number of participants when using medication-conditions list in four resources

Medication	**Conditions treated by medication**	**Participant num**	**Condition completeness**
Acetaminophen	Fever;Headache;Toothache;Degenerative polyarthritis; Myalgia ;Hyperthermia; Migraine Disorders; Common Cold; Back Pain; Pharyngitis; Arthritis; Dysmenorrhea; Influenza;Pain;	165794	0.729
lidocaine	Poisoning; Pain finding; Unspecified disorder of esophagus; Coronary arteriosclerosis in patient with history of previous myocardial infarction; Ventricular tachycardia, unspecified;Other type of myocardial infarction;	124291	0.504
Sodium chloride	Adiposity; Dry eye; Other obesity not elsewhere classified; Tear film insufficiency, unspecified;Overweight and obesity; Tear film insufficiency;Sodium deficiency;Dehydration;Genetic susceptibility to obesity	123176	0.536
Ondansetron	Alcoholic Intoxication, Chronic;Nausea;Vomiting;Postoperative Nausea	112985	0.363
fentanyl	Pain, Postoperative;Pain, not elsewhere classified;Observation of pattern of pain;Discomforting present pain;Chronic pain;Pain, unspecified;Acute onset pain;Pain finding;Pain	96776	0.569
oxycodone	Pain, Postoperative;Chronic pain;Severe pain;Pain;Postoperative pain	89375	0.594

ibuprofen	Rheumatoid bursitis, unspecified hip;Ankylosing spondylitis of cervicothoracic region;Fever;Chronic gout, unspecified;;Rheumatoid Arthritis;Ankylosing spondylitis;Common Cold; Migraine Disorders;Hyperthermia;Dysmenorrhea (finding);Pain, Postoperative;	84651	0.815
midazolam	Agitated ;Absence epileptic syndrome, intractable, with status epilepticus;Unspecified dementia;Dravet syndrome, intractable, with status epilepticus;Anxiety disorder; Lennox-Gastaut syndrome, intractable, with status epilepticus; Psychomotor agitation;	78495	0.373
albuterol	Asthma;Bronchial Spasm;Bronchospasm; Chronic Obstructive Airway Disease;Pneumopathy;Lung diseases;	71425	0.823
propofol	Dementia;Epileptic;Agitation;Delirium caused by substance or medication;Lennox-Gastaut syndrome, with status epilepticus; Pain;Organic brain syndrome;Lafora progressive myoclonus epilepsy, not intractable, with status epilepticus;Psychomotor agitation;	69251	0.232
aspirin	Rheumatoid bursitis,Fever ;Rheumatoid arthritis; Cerebrovascular accident ; Inflammation; Pharyngitis; Rheumatoid arthritis; Degenerative polyarthritis; Streptococcal Infections; Observation of pattern of pain; Myocardial Infarction; Angina, Unstable; Ankylosing spondylitis; Pre-Eclampsia; Coughing; Gout; Headache; Transient Ischemic Attack ;Mucocutaneous Lymph Node Syndrome; Heart Diseases;	65326	0.898
docusate	Constipation	64307	0.366

References

1. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2016;
2. Weiskopf NG, Hripcsak G, Swaminathan S, Weng C. Defining and measuring completeness of electronic health records for secondary use. Journal of biomedical informatics. 2013 Oct 1;46(5):830-6.
3. VHA V. National Drug File Reference Terminology (NDF-RT) Documentation. US Department of Veterans Affairs. 2012.
4. Smith JC. Adverse Drug Effect Detection for Clinical Decision Support (Doctoral dissertation).