Identification of HIV positive individuals across multiple datasets

Craig S Mayer, MS[1]
[1]Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD

Background

Attributing medical conditions to individuals can be done in a variety of different ways, including through diagnosis, lab measurement, clinical observation, or self-reporting.[1,2] In the case of data reuse, condition attribution may be dependent on the amount and type of data being analyzed. Understanding what type of data is being used is key in understanding how best to maximize patient capture for different medical conditions. The use of harmonized data aids in performing condition attribution by allowing for applying a single value set to multiple datasets and quick comparison across dataset types. However, the impact of source data differences is under described on value set capture of distinct patients. The objective of this study was to attribute HIV positive status to individuals across multiple OMOP CDM transformed datasets through a variety of methods. The analysis was also done to showcase the differences in how a condition can be attributed based on the type and breadth of the dataset being used.

The analysis included three datasets 1) the All of Us (AoU) program, 2) UK Biobank (UKBB), and 3) Clinical Practice Research Datalink (CPRD) AURUM. AoU is a large United States based data collection initiative that combines survey with integrated electronic health record (EHR) data.[3] UKBB is a large health and biomedical database based in the United Kingdom which includes a combination of collected questionnaire, biological sample and integrated inpatient EHR data.[4] CPRD AURUM is a data collection initiative that collects data from a set of United Kingdom based general practitioners (GPs).[5,6]

Methods

Using OMOP versions of AoU, CPRD and UKBB and a set of HIV related OMOP concept_ids, HIV positive individuals were identified through multiple methods and distinct patient counts from each method were calculated.[7] This included condition, measurement, and observation domains. For condition, an individual was considered HIV positive if one diagnosis code for HIV was present at any time or context. For measurement the presence of a single viral load test or a positive confirmatory test (such as Western Blot) with previously defined value cutoffs were used for HIV status attribution.[8] The exclusion of screening tests may lead to missing individuals, but were excluded due to the possibility of false positives and the normal practice of a follow-up confirmatory test before real world HIV attribution. The observation domain included indication of self-reported HIV positive status through survey responses regarding personal medical history. Due to the chronic nature of HIV, an indication of history of HIV was deemed sufficient in considering an individual as HIV positive. For AoU and UKBB each domain (condition, measurement, and observation) was included, while CPRD only included condition and measurement, as self-reported medical history was not separated in the source data. A comprehensive code list can be found at the project repository.[9] The populations from each domain were compared within each dataset to determine any crossover and exclusion between each attribution method.

Results

AoU includes 413,457 total individuals, while UKBB includes 502,390 and CPRD includes 49,102,289. For UKBB only 9,689 (1.93%) had lab measurements done on biological samples provided. There were 7,337 distinct HIV cases (1.77% of total AoU population) in AoU, while there were 484 distinct (0.10%) cases in

UKBB and 50,374 distinct (0.10%) in CPRD. In total, for the three datasets, 58,192 HIV cases were found. Table 1 shows the case distribution by domain.

Table 1. HIV positive case count by domain.

Domain	AoU	UKBB	CPRD
Condition	5,185	214	50,374
Observation/Self-reported	1,686	484	X
Measurement	3,925	18	2,806

In many cases individuals were present in multiple domains. Table 2 shows the crossover between the different domains.

Table 2. Case count crossover into multiple domains

Domains	AoU	UKBB	CPRD
Condition and Observation	1,031	194	x
Condition and Measurement	2,403	18	2,806
Measurement and Observation	575	18	x
Condition, Measurement and Observation	550	18	x

In Table 2, every identified HIV positive individual in CPRD who was identified via measurement was also present in condition. This means the condition attribution in CPRD captured all cases in the dataset. The fact that no positive HIV measurement cases were exclusively found (outside the condition cases) indicates a lack of false positive lab tests in the CPRD analysis. All individuals who tested positive were indicated to be HIV positive through a diagnosis from the medical provider. For UKBB, all HIV positive cases in the condition and lab measurements were in the observation (self-reported). It is difficult to assess attribution inclusivity of UKBB measurement due to the limited number of participants who were tested (1.93%). For AoU each case capturing method included unique individuals absent in other attribution methods. Each attribution method in AoU yielded additional HIV positive cases, as shown by observing that only including diagnostic data would have excluded 29.33% of HIV positive individuals found, as they were found in other domains. The results also indicate the importance collecting medical history rather than just being reliant on EHR importation. For AoU and UKBB, removing self-reported medical history would remove 2,170 (27.75% of all AoU and UKBB cases) HIV positive cases.

Conclusion

Distinct patient capture for a given condition is dependent on the source data being analyzed. This is exemplified by the fact that all individuals in the studied primary care (CPRD) data were included via

the condition domain. In this case, since the data captures primary care data, all positive HIV measurement individuals were captured via diagnosis as they are receiving care via the GP data. Conversely for a data collection program (AoU and UKBB), the results showed the utility of combining multiple domains (condition, measurement, and observation) to identify all individuals with a given condition. In the case of AoU each domain yielded additional HIV positive cases that otherwise would have been missed. While this analysis was limited to HIV and these select datasets, the principles and analysis are generalizable to any condition or dataset and can be reused in the attribution of conditions in other OMOP CDM datasets. Due to the nature of different conditions and how they are normally attributed the identification of individuals for a given cohort may vary based on the type of data included as shown by our results and the exclusion of a domain for attribution may limit the accuracy of condition attribution. These factors should be considered in defining cohorts and data selection.

Acknowledgement

This work was supported in part by the Lister Hill National Center for Biomedical Communications of the National Library of Medicine (NLM), and in part by the Office of AIDS Research (OAR), National Institutes of Health. I would like to thank James Mork and Nick Williams for comments on drafts of this report.

References

1. Evans HE, Tsourapas A, Mercer CH, Rait G, Bryan S, Hamill M, et al. Primary care consultations and costs among HIV-positive individuals in UK primary care 1995-2005: a cohort study. Sex Transm Infect. 2009;85(7):543–9.

2. Evans HE, Mercer CH, Rait G, Hamill M, Delpech V, Hughes G, et al. Trends in HIV testing and recording of HIV status in the UK primary care setting: a retrospective cohort study 1995-2005. Sex Transm Infect. 2009;85(7):520–6.

3. All of Us Research Program Investigators, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, et al. The “All of Us” Research Program. N Engl J Med. 2019 Aug 15;381(7):668–76.

4. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med. 2015 Mar 31;12(3):e1001779.

5. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, et al. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol. 2015 Jun;44(3):827–36.

6. Lee G. CPRD Aurum Data Specification. :14.

7. Papez V, Moinat M, Voss EA, Bazakou S, Van Winzum A, Peviani A, et al. Transforming and evaluating the UK Biobank to the OMOP Common Data Model for COVID-19 research and beyond. Journal of the American Medical Informatics Association. 2022 Oct 13;ocac203.

8. Kranz LM, Gärtner B, Michel A, Pawlita M, Waterboer T, Brenner N. Development and validation of HIV-1 Multiplex Serology. Journal of Immunological Methods. 2019 Mar;466:47–51.

9. CRI/HIV at master · lhncbc/CRI · GitHub [Internet]. [cited 2023 Aug 23]. Available from: https://github.com/lhncbc/CRI/tree/master/HIV