Validation of Real World Data: Case study in Hepatitis C

Craig Mayer[1], Nick Williams[1],Sigfried Gold[1], Kin Wah Fung[1], Vojtech Huser[1]
[1]Lister Hill National Center for Biomedical Communications, NLM, NIH, Bethesda, MD

Background

Real World Data is continually being used to generate clinical evidence and make observations regarding patient disease diagnoses. The development of patient phenotypes and the validation of accurate diagnoses is important in confirming inclusion in a cohort and assessing patient outcomes. Validation of such phenotype algorithms is important for using Real World Data (RWD) to generate Real World Evidence (RWE) from a cohort of patients. With this in mind, our study looks at two of the major type's of phenotypes, diagnosis coding and lab measurements, and compare the overlap and inclusivity of each phenotype to validate the correctness of the included Real World Data.

A common coinfection in HIV positive patients is Hepatitis C. Hepatitis C can be diagnosed via a lab test for HCV RNA Viral Load.

Methods

Two cohorts of HIV-infected patients were analyzed for Hepatitis C coinfection.
The first cohort was from Electronic Health Records (EHR) data obtained through collaboration with the Greater Plains Collaborative (GPC). The second cohort was generated from the NIH Clinical Center research data warehouse (called BTRIS).

We looked at the cohorts from GPC and BTRIS and created two phenotypes for comparison of Hepatitis C coinfection from each data source. We developed a diagnosis (DX) phenotype by taking a set of ICD-9CM and ICD-10CM codes from each cohort that represent a Hepatitis C diagnosis and included each patient that had at least one of these diagnosis codes appear at least once.

We also created a measurement cohort (MS) which was comprised of lab measurement data for HCV RNA viral load. In the GPC data, this was represented by LOINC code 20416-4. In the BTRIS data, this was represented by test name and so a string search for Hepatitis C viral load was done. We included every patient with at least one positive or detectable Hepatitis C lab test.

Finally, considering diagnosis the standard for GPC and measurement the standard for BTRIS, we evaluated sensitivity, specificity and PPV of the other cohort. We compared the differences in the composition of the cohorts and looked at the overlap between the two phenotypes..

Limitations: We only tested two phenotypes. We also only included diagnosis codes specific to Hepatitis C and excluded any unspecified hepatitis or unspecified infectious diseases which would have included more actual positives as well as significantly more false positives. We also limited our measurement cohort to just those that received a positive HCV viral load test and not any other potential diagnostic tests. We also based cohort inclusion based on just one diagnosis code or one positive lab test. Further work would include the inclusion of a third phenotype, namely medication data, which would further verify true positives and show better phenotype differentiation.

Acknowledgements: This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH)/ National Library of Medicine (NLM)/ Lister Hill National Center for Biomedical Communications (LHNCBC) and by the NIH Office of AIDS Research. The findings and conclusions in this article are those of the authors and do not necessarily represent the official position of NLM, NIH, or the Department of Health and Human Services. **Contact:** craig.mayer2@nih.gov. ICF Inc. provided contractual support for carrying out analysis on this project.

Results

In GPC we found a total of 252 patients that either had a Hepatitis C diagnosis code or a positive HCV viral load test. 50.4% had both a diagnosis code and a positive lab test, 42.8% had just a diagnosis code and 6.8% had just a positive lab test

In BTRIS we found 744 participants that qualified for at least one of the phenotypes. 59.9% were in both, 7.1% just had a diagnosis code and 32.9% just had a positive lab test.

We used the measurement driven phenotype algorithm as silver standard to determine sensitivity and specificity (see Table 1). As can be seem by these results a majority of patients with Hepatitis C meet the criteria of both a diagnosis code and a positive lab test. However the difference between the two cohorts is that GPC has a much larger percentage of patients with just a diagnosis code while BTRIS had a much larger percentage with just a positive lab test.

Table 1: Results comparing GPC and BTRIS datasets

Dataset	Greater Plains Collaborative)	BTRIS
Type	EHR	research data warehouse
Total Patients(n)	2033	4040
Diagnosis Code Identified (DX)	235	499
Measurement Identified(MS)	144	691
In both DX and MS	127	446
In just DX	108	53
In just MS	17	245
Specificity	54.0%	64.5%
Sensitivity	98.8%	98.5%
PPV	88.2%	89.6%

Conclusions

We found that in our analysis, our EHR data (GPC) is more representative based on diagnosis code than on lab codes in diagnosing Hepatitis C patients as more patients were captured using diagnosis codes. This is possibly explainable as our study only includes a segment of time that does not include a patient's full life cycle so a positive test could exist earlier for the patient then the records analyzed, while a diagnosis code will remain with a patient usually through their record life cycle. This also pertains to patients who may have joined the cohort (EHR) after a positive lab test and diagnosis were confirmed previously at another institution or lab measurements were done outside the EHRs range.

For clinical research (BTRIS) we found the opposite to be true for our data, as the measurement data is more encompassing than diagnosis codes. This is likely due to the fact that in a research setting lab tests are commonly used to verify diagnosis and confirm enrollment in a study as well as the fact there is more standardized and consistent lab testing for patients enrolled in a research study. A related OHDSI R package PheValuator is addressing a similar problem; it was announced after our work had already been initiated.

U.S. National Library of Medicine