PharmaSUG 2021 - RW-126

A Programmer's Experience Researching Real-World Evidence (RWE) COVID-19 Data

Ginger Barlow, UBC

ABSTRACT

In early 2020, when the COVID-19 pandemic hit the US, researchers started gathering and analyzing data to help understand who was vulnerable to the infection, make major policy decisions to reduce the spread of the disease and start the race for an effective vaccine. At UBC, a team of researchers used the Electronic Health Record (EHR) data from Cerner Health Systems' Real-World Data™ (CRWD) to analyze records of 14,371 inpatients with a COVID-19 diagnosis. The primary objective of the study was to use real-time data from the CRWD to assess how variation in coding algorithms to classify cases of COVID-19 impacted the number and types of patients identified.

Our statistical programming team was called to analyze the data but there were many challenges in doing this type of research including how different the data in the RWE database was compared to interventional or observational trials data; classifying patients into 4 cohorts based on how their diagnosis was made possibly across multiple encounters; and the lack of universal accurate laboratory testing and heterogeneity in ICD-10 codes. We defined 4 mutually exclusive COVID cohorts and then compared comorbidities, medications, complications, and length of stay. The programming challenges were interesting, and the conclusions of the study highlighted how important case definitions and algorithms are when utilizing EHR data for research.

INTRODUCTION

This paper will describe how the programming team used EHR data from CERNER data to build reports for a publication, the challenges we uncovered and how we worked closely with the researchers to build algorithms that defined cohorts accurately for analysis. Reviewing a case study like this can benefit programmers who may be new to working with real time, real world data by highlighting some of the situations they may encounter.

THE DATA

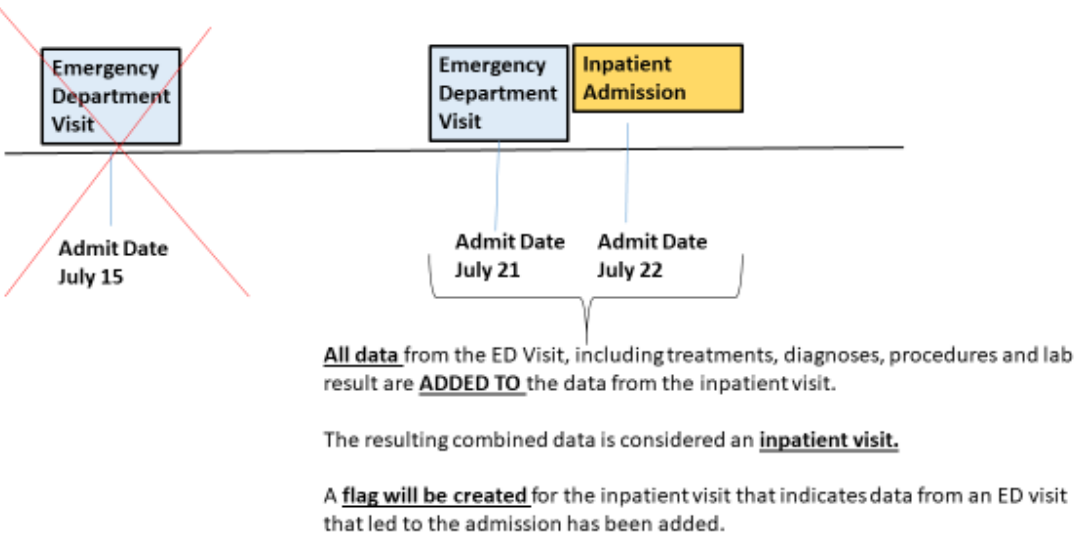
Data from CRWD was delivered to UBC in CSV files. The data was in flattened structure with 7 tables with an average of 10 variables each except for the results file that had 30 variables. The key variables in the datasets were ENCOUNTERID and PERSONID. ENCOUNTERID was unique for each ER visit, urgent care visit and hospitalization.

The team had provided criteria for data to be pulled from CERNER. After examination of the data, we determined that we could not use all the patients' data provided because they did not meet the eligible population definition. We had to programmatically identify eligible patients and which encounters would contribute data for the analysis. We also had to re-derive flags due to changing specifications as data was analyzed.

DETERMINING ELIGIBLE CASES

First, we had to identify which encounter would be the primary one to use for analysis and use that index date to select eligible cases for our study. This is shown in Figure 1.

Error! Reference source not found. Identifying Index Encounter



In the scenario above, the patient's index date was derived as July 22nd and data collected from both the emergency department visit immediately before admission and the inpatient admission would be included in analysis. If a patient had multiple inpatient admissions, the last admission would be used. A patient could only contribute one hospitalization.

Once this index encounter was identified, we selected all the cases (patients) that were eligible for the study. The eligible population was defined as patients with an in-patient hospital stay who were one of the following:

- Laboratory confirmed COVID-19 positive, that is having a positive lab test during the hospital stay or within two weeks prior to the hospitalization
- Clinically diagnosed with COVID-19 (using ICD-10 codes) without positive COVID-19 laboratory results during the stay or within two weeks prior to the hospitalization
- Diagnosed with suspected exposure during the encounter or 2 weeks prior

If the index encounter met any of these criteria, the patient was considered within the eligible population and that ENCOUNTER_ID was used to select data for analysis. Of the 16,900 patients who were determined by CERNER as qualified for the COVID study population, 14,371 patients were selected as being eligible.

REAL-TIME DATA CHALLENGES

There were some unique challenges we encountered due to the real-time nature of the data.

Laboratory codes (LOINC) were provided to identify COVID-19 tests but with the novelty of the virus, in the early data, codes were not available for specific COVID-19 testing. The clinicians in the research team had to review data and determine which codes would be included as COVID-19 tests. The selected codes included tests for human coronavirus, SARS coronavirus and SARS-related coronavirus. These files were provided to programming in CSV format to import and combine with the CERNER data.

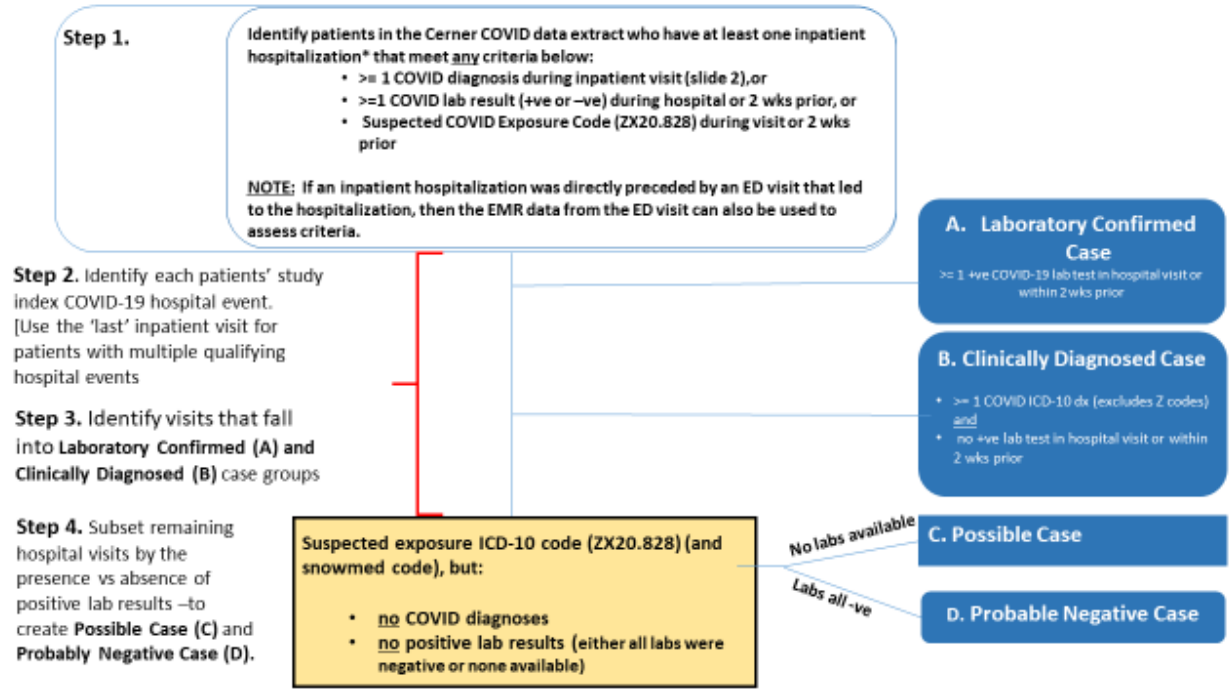
Identifying diagnoses codes was also challenging because new ICD-10 codes were being created during the period of our study. In April 2020, the CDC issued new coding and reporting guidelines related to COVID-19, so data collected before that time reported diagnoses using various, less-specific codes. Since then, the CDC has issued new guidelines including six new codes effective January 1, 2021. Patients being admitted to the hospital with symptoms of COVID-19 but not having positive lab tests might have been diagnosed with bronchitis, lower respiratory infection, or even just potential exposure to COVID-19. Again, the clinical research team reviewed ICD-10 codes from the data and coding dictionaries and determined which would be used to classify clinically diagnosed cases and provided files to programming to import and combine with the CERNER data.

CASES TO COHORTS

The Statistical Analysis Plan (SAP) described how cases would be divided into 4 separate, mutually exclusive cohorts as shown in Figure 2 and determined in this order:

1. Lab-confirmed COVID-19 hospitalization cohort:
    a. Patients who have at least one positive COVID-19 lab test result at any time during the hospital stay or within 2 weeks of admission
2. Clinically diagnosed COVID-19 hospitalization cohort - Patients who do not qualify for the lab-confirmed cohort but have A diagnosis consistent with COVID-19 infection (using ICD-10 codes) and one of the following:
    a. At least one negative COVID-19 diagnostic test during the index encounter or 2 weeks prior and no positive tests during that same time period
    b. No evidence of a COVID-19 lab test during the hospital stay or in the 2 weeks prior to admission
3. Possible COVID-19 hospitalization cohort – Patients who do not qualify for either of the cohorts above and have a diagnosis consistent with the COVID exposure code (Z20.828) and no evidence of:
    a. Any COVID-19 test results (positive or negative) during the index encounter or 2 weeks prior
    b. Any COVID-19 diagnosis during the hospital visit
4. Probable negative COVID-19 case – Patients who do not qualify for any of the cohorts above who:
    a. Have a diagnosis code of suspected or possible COVID-19 diagnoses
    b. Have at least one lab test for COVID-19 during the hospitalization or two weeks prior
    c. Have all negative lab results

Figure 2 Cohort Selection



FINALLY, THE ANALYSIS

Our study now had 14,371 cases to analyze, with 6,623 (46.1%) being lab-confirmed (positive COVID-19 test in the relevant time period) and 7,748 (53.9%) cases clinically diagnosed without positive lab tests.

To produce the report, data needed to be collected from other files using the index ENCOUNTERID for demographics, comorbidities, procedures and complications and other data of interest. When reviewing the data, we again found scenarios in data that needed to be discussed and addressed to develop algorithms for which data to analyze.

Initially, we were planning to only analyze data collected at the index hospital admission or an emergency department visit leading to that admission. However, we found cases where the COVID diagnosis had been made at encounters prior to the derived index encounter date but not reported at the index encounter. Some patients were diagnosed with COVID before going in for an ER visit or being admitted to the hospital, so that diagnosis was not linked to the ENCOUNTERID we were using as the index encounter. The team had to agree on a new algorithm to include COVID-19 diagnoses made at any encounter prior to or including the index visit.

In the original data received, a flag had been derived to indicate a positive lab result at the encounter visit or within 14 days prior. However, we found this flag did not include the LOINC codes that had been identified to include as COVID-19 lab tests and considered results outside of the 14-day period prior to admission. We had to derive that flag using our data rather than using the flag that had been provided.

In the SAP, we hadn't planned to have analysis of the duration of hospital stay but when the team decided to do that, we ran into additional data challenges. While hospital admission dates were recorded, discharge dates might not be because the data was real time and patients might still be in the hospital. In the case of deaths, the death date was included in the data but not as a discharge date. We also needed to examine those cases where an ER or urgent care visit lead to hospital admission but in some cases, we found the ER date was more than a day before hospital admission which is not typical. After review of those types of scenarios, the researches manually reviewed some cases and provided a list of additional ENCOUNTERIDs to use to select data from.

When viewing comorbidity tables, the researchers found that there were very low instances of common conditions such as hypertension or diabetes. With further data review, we found that some comorbidities might not be reported at all if diagnosed before EHR data was transferred into the CERNER database. Or, if the information was in the database, it may have been collected at various times other than the index encounter including in other health records, previous urgent care visits, or prior hospitalizations. This was especially prevalent with the long-term chronic conditions that are directly related to COVID-19 outcomes. We had to change our algorithms to look across all encounters rather than just the index encounter. When the numbers still didn't reflect the incidence of those diseases in the general population, the medical team decided to compile a list of medications to use to flag patients with certain comorbidities when there was no diagnosis code reported. One example was to flag a patient as diabetic if they were on insulin but there was no diagnosis of diabetes.

It seemed with every program we worked on, more situations arose that needed discussion, new algorithms, and additional programming.

CONCLUSION

With all these reviews, meetings, and data discussions behind us the team was ready to produce tables for the publication. After over 100 person hours of programming, we produced 6 tables. Yes, only 6. We had tables for demographics, baseline characteristics, concomitant medications, comorbidities, treatments, and procedures.

Remember, the primary objective of this study was to use EHR real time data to evaluate the variability in COVID-19 case capture and implications on epidemiologic research. We learned just how important it can be to evaluate the algorithms that identify cases and cohorts to see if they match the intent of the analysis. For example, if we had only looked at comorbidities identified in that dataset, we would have missed many chronic conditions like diabetes and hypertension. By considering medications as well, we more closely matched the known percentages of chronic disease in the general population.

When using real time, real world data, a programmer certainly cannot "just follow the specs." The programmer must be actively involved in researching data, bringing anomalies or unexpected situations up for discussion and working with the team to come up with solutions to ensure the resulting analysis is reflective of the statistical plan. Working on a project like this is an exciting foray for programmers to interact directly with researchers and be a strong contributor to a research project.

REFERENCES

CDC. 2020. "ICD-10-CM Official Coding and Reporting Guidelines April 1, 2020 through September 30, 2020." Accessed April 2, 2021.

https://www.cdc.gov/nchs/data/icd/COVID-19-guidelines-final.pdf

ACKNOWLEDGMENTS

The author would like to recognize members of the UBC research team who she had the pleasure of working with on this exciting research: Irene Cosmatos, Micheal Bulgrien, Scott Horton and Ching Tu. This research project was done in conjunction with researchers from Duke University.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ginger Barlow
UBC
215-260-9375
ginger.barlow@ubc.com