Longitudinal Analysis of COVID-19 Outcomes in the US Population

Topic: Long-haul COVID-19 complications

Jacelyn Dorman, u6032430@utah.edu, u6032430 | Clinical SME

Matt Hoffman, Matthew.Hoffman@hsc.utah.edu, u0151460 | Data Science SME

Nikhil Kala, nikhil.kala@utah.edu, u1319777 | Technical SME

Kaitlyn Stevens, u1025471@utah.edu, u1025471 | Clinical SME

Joshua Uda, joshua.uda@utah.edu, u1377747 | Business Analysis & Project Management SME

Github: https://github.com/nikhilkala/Data-Wrangling-Covid-Team4

Background and Motivation

The COVID-19 pandemic has had a profound impact on the world, causing widespread illness, death, and disruptions to daily life. While much has been learned about the acute effects of the virus, there is still much unknown about its long-term effects. Studying the long-term effects of COVID-19 is crucial for understanding the full scope of the disease and the ongoing impact it may have on individuals who have recovered from acute infection. This knowledge can inform public health policy, guide medical treatment and support, and ultimately help improve health outcomes for those affected by the virus. Additionally, studying the long-term effects of COVID-19 will be important for preparing for future pandemics and improving our ability to respond to similar public health crises in the future. Furthermore, this study will drive the quintuple aim of healthcare improvement by improving population health, enhancing patient experience, reducing costs, improving provider satisfaction, and advancing knowledge.

Project Objectives

The primary objectives of this study are to:

1. Identify trend anomalies in conditions that may be associated with long-COVID-19 syndrome
2. Report on the statistical significance of correlation between COVID-19 and concomitant conditions
3. Recommend focus areas for additional research to evaluate potential COVID-19 etiology

It is our hope that the results of the study can inform public health policy and support

decision-making in the allocation of resources and the development of interventions aimed at reducing the impact of the virus, provide valuable information for healthcare providers and organizations so they will be better equipped to develop differential diagnoses, medical treatment plans, address the needs of patients, and provide ongoing support to patients who have recovered from acute COVID-19 infection, and contribute to the body of knowledge on the long-term effects of COVID-19, which will be important for preparing for the long-term clinical, operational, and financial impacts of the pandemic. Some research questions we intent to explore are:

- Is COVID-19 infection associated with onset of rare conditions such as auto-immune disorders?
- Can time-series anomalies be detected for any conditions over the time period of the pandemic?
- Are any conditions statistically correlated, and not just concomitant, with COVID-19 infections?
- Which anomalous signals are most prevalent, e.g., labs, diagnosis, procedure, prescription trends?

Data

We are using the TriNetX Covid-19 research data set with 73 million patients and 25 billion facts. The dataset is found here: https://uofu.box.com/s/pip0k5ky86g9qayl85y0j8uwub13p7tw.

This dataset contains:

- Values from structured EHR fields (e.g. demographics; date-indexed encounters, diagnoses)
- Facts and narratives from free text (e.g. medications identified through NLP)
- Death dates from mortality registries
- Tumor morphology and size data from tumor registries and surgical pathology reports

The file format is CSV and is coded in UMLS standards.

Source and Master Terminologies by Data Domain

Domain	Example Source Terminologies	Target TriNetX Terminologies
Demographics	Various, uncoded, HL7	HL7
Diagnosis	ICD-9-CM, SNOMED CT, ICD-10-CM, ICD-10	ICD-10-CM
Procedure	Non-standard terminologies unique to organization, ICD-9-CM, SNOMED CT, ICD-10-PCS, CPT, HCPCS	ICD-10-PCS, CPT, HCPCS, SNOMED CT
Medication	WHODrug, NDC, RxNorm	RxNorm, with OMOP extension for medications not approved in the U.S.
Lab	Non-standard terminologies unique to lab, SNOMED CT, LOINC	LOINC
Oncology	Various	ICD-O
Genomics	Various	HGVS
Vital Signs	Various	LOINC
Visit Types	Various	HL7

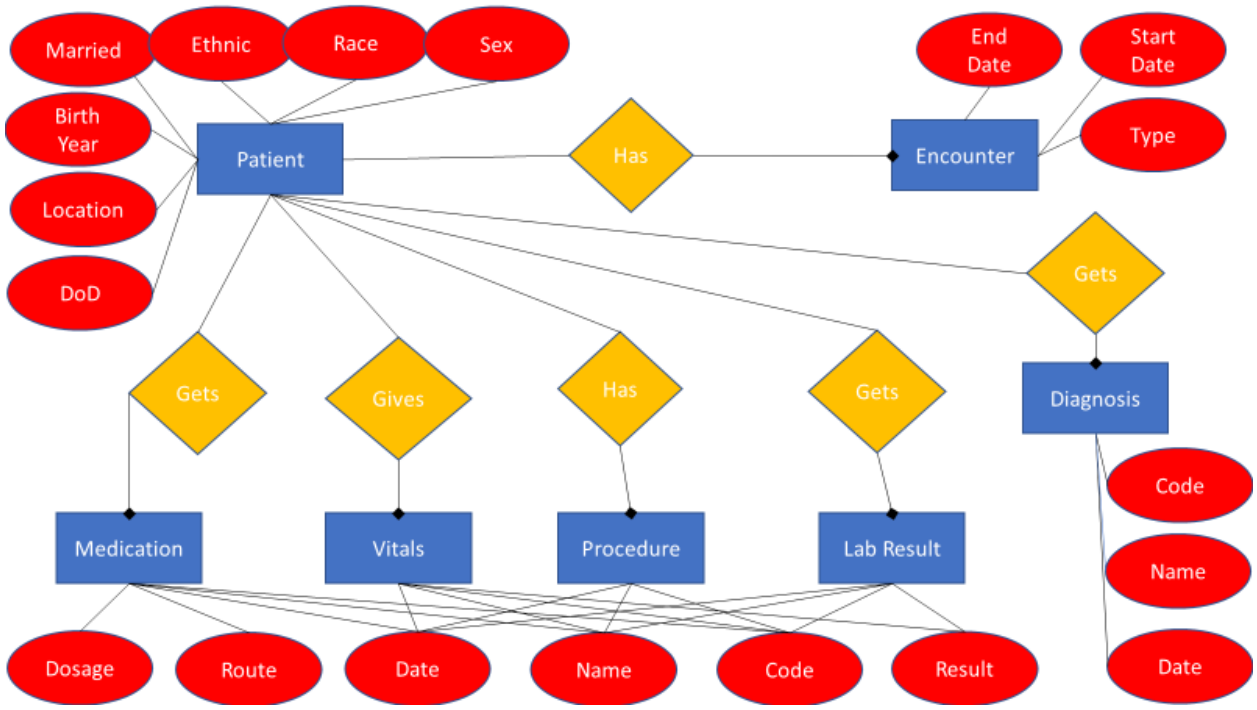
Data Processing

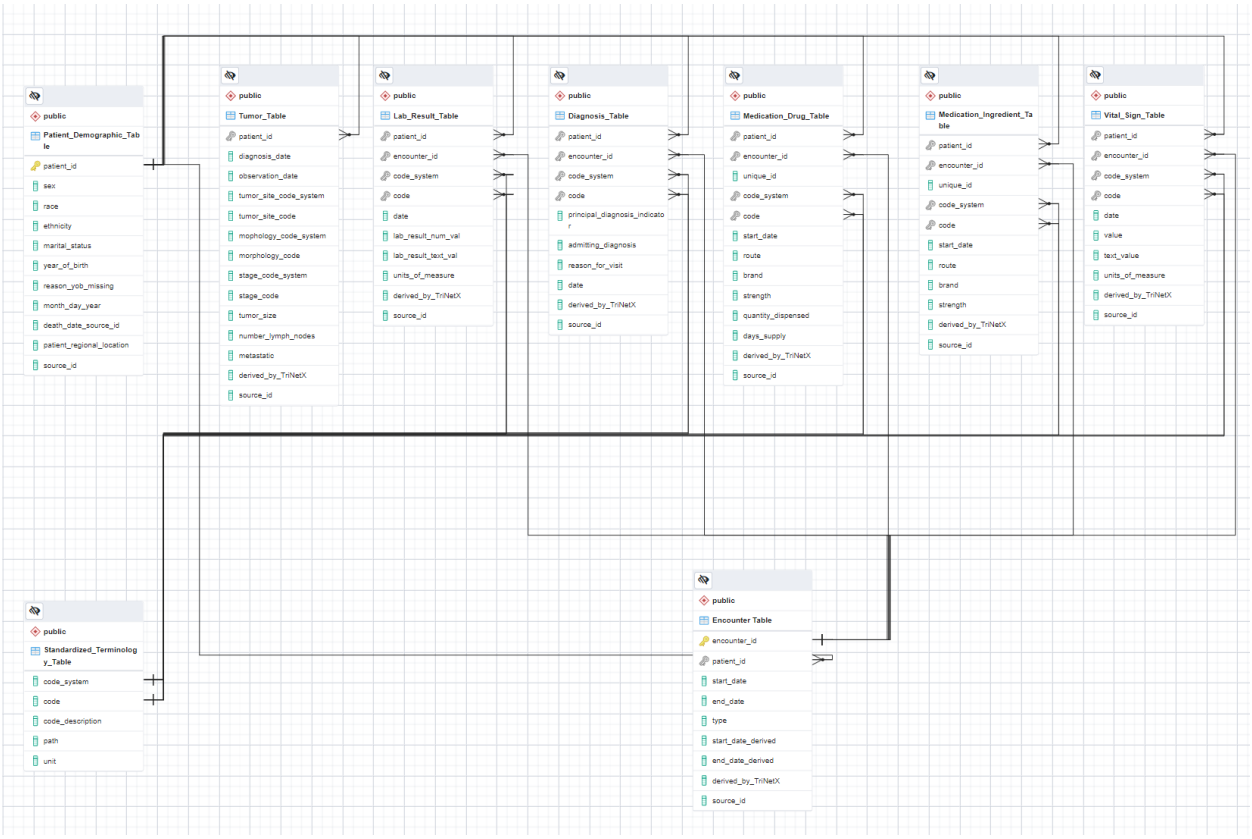
The dataset involves many sources in terms of the CSVs. Data inconsistency, redundancy, missing data, and other quality aspects will be assessed. To overcome quality issues we do data cleanup. This will include tasks such as data normalization, removing empty values, and outlier detection. Since the dataset is expansive we will not use Excel, but instead will be using Python for the analysis. Python has a rich ecosystem of libraries for data processing and analysis. We will be using Python, Jupyter, Pandas, NumPy, Scikit, and Openpyxl. The dataset is huge with lots of data points so we will prune the data according to our needs. This will also be helpful as we can choose data which has no quality issues, so the cleanup will be easier.

The data processing will be done with Pandas. Using Pandas, all the CSVs will be imported as Dataframes. These data frames are each representative of an individual excel file. Using the needed features we will drop the columns which are not needed and check in the subsequent rows to see if any values are empty or are invalid in nature. The dataframes can then be made into a single dataframe, which represents our entire data together. Then the data can be normalized and do other cleanups if needed. Finally, that dataframe can be exported and saved as CSV, and can in future be imported directly to analyze the data for features.

Design

1. Relational model outlining how the various tables and data sources are connected:





2. Data completeness analysis
    a. Chart showing percentage of complete records compared to incomplete records
    b. For incomplete records a chart showing breakdown of offending data points compared with each other and the complete record set
3. Continuous data quality dashboard
    a. Statistical analyses
    b. Box Plot analyses
4. Discrete data dashboard
    a. Histograms of each of the discrete data points showing the distribution
5. Categorical data dashboard
    a. Bar chart of the counts of the different categories

Must-Have Features

- Diagnoses
    - Diagnosis code(s)
    - Diagnosis date(s)
    - Treatment/procedure(s)

        - Medication(s)
- Patient Demographics
    - -Age
    - -Sex
    - -Race
    - -Death Date (if applicable) and reason
    - -Regional Location
- COVID-19 information
    - -diagnoses date
    - -treatment

This dataset is going to be used for identifying signal anomalies in potential COVID-19 induced conditions. With that being said, there are some features of the data that are absolutely necessary in order for this project to be a success. The first crucial information is regarding the patient demographics that describe the patient; this includes age, sex, race, their regional location, and their death date (if applicable). Along with the date of death, the reason for death will provide useful information about the severity of diagnosed disease. Diagnosis is also very important for this is the data that the analysis will most heavily rely on. For this, it's crucial to know the diagnoses, as well as the categorical diagnosis codes, the date of diagnosis, what treatments or procedures were done for the diagnosis, and if the patient is taking any medication for the diagnosis. These diagnostic features can help detect for signal anomalies in conditions that may have been induced by COVID-19, while the demographic features may be able to help in detecting other patterns Because this study is concerning conditions that may have been triggered by COVID-19, it is important that there be information about when the patient was diagnosed with COVID-19 and what was done for treatment for it. It's also important to know the timeframe between a COVID-19 diagnosis and future diagnosis that are being explored, but this can be calculated using the other data.

Optional Features

- Diagnosis codes all having the same format (ICD-10, SNOMED, etc.)
- Social Determinant of Health Risk
- Pre-existing conditions/comorbidities
- Vaccination status

While there are some features that are crucial for the dataset to have in order for the study and analysis to be a success, there are other features that could provide additional insights and be beneficial, but are not necessary. First, it would be ideal if the dataset had the diagnosis codes all use the same classification for the sake of simplicity and possibly to perform less data wrangling. Next, it could provide more insight if there was more information provided about pre-existing conditions and comorbidities that the patient has been diagnosed with. While the analysis is focused on the conditions and diagnoses prior to COVID-19, this additional data could potentially highlight patterns between these conditions, COVID-19, and future conditions. With that, it could also be useful to know the vaccination status of the patient, which could show if there is a correlation

between future diagnoses and vaccination status. The last additional information that would be helpful, but isn't necessary for this analysis is information to determine a patient's Social Determinant of Health Risk (SDOH). This could potentially help reduce the risk of a confounding variable and highlight a different aspect affecting a person's health, that isn't COVID-19. None of the previously mentioned variables are necessary, but they could provide additional information to strengthen the analysis of the data.

Project Schedule

- [x] ~~January 26 - Charter Team~~
- [x] ~~February 6 - Complete Proposal Rough Draft~~
- [x] ~~February 11 - Select Dataset~~
- [x] ~~February 13 - Complete Proposal & Review~~
- [x] ~~February 16 - Submit Project Proposal~~
- [x] ~~February 20 - Complete Data Exploration~~
- [x] ~~February 27 - Complete Data Quality Assessment~~
- [x] ~~March 2 - Instructor Inspect-Adapt 1~~
- [ ] March 6 - Complete Data Cleaning
- [ ] March 13 - Complete Data Diagrams
- [x] ~~March 16 - Update Github Submission~~
- [ ] March 20 - Complete Schema Design & Queries
- [ ] March 22 - Intermediate Presentation & Review
- [ ] March 27 - Complete Data Dashboards
- [ ] April 3 - Complete Draft Presentation
- [ ] April 6 - Instructor Inspect-Adapt 2
- [ ] April 10 - Complete Final Presentation
- [ ] April 17 - Rehearse Final Presentation
- [ ] April 24 - Record Final Presentation
- [ ] April 27 - Final Presentation & Review

- [ ] May 4 - Final Project Submission