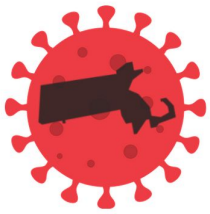
Visualizing synthetic COVID-19 data in Massachusetts
CSE 583 - Functional Specification
November 20, 2020
Aja Sutton, Andrew Teng, Jason Thomas, Nanhsun Yuan



Background:
The novel coronavirus (COVID-19) has dramatically altered how we interact and socialize with one another. Our lives have shifted to balancing public health guidance and policies aimed at reducing the spread of the virus, with attempts at maintaining a sense of normalcy. Current COVID-19 research focuses on increasing our understanding of how the virus spreads through communities and neighborhoods. Given the novelty of the virus, researchers face many challenges and unknowns. First, obtaining access to patient data can involve a lengthy bureaucratic process, especially as patient data is protected and governed by the Health Insurance Portability and Accountability Act (HIPAA). Second, understanding the trends of the data may be difficult as data representation and visualization methods are highly variable, making the data subject to interpretability. Synthetic data -- simulated data that are generated based on the trends and patterns of real data -- may provide an avenue for researchers to better understand real-world data trends without the need to overcome the obstacles involved in obtaining real patient data. Being that synthetic data may be modeled on real-world data, it may allow researchers to generate results that are remaining meaningful and translatable while being easily accessible.

Goals:
- With the synthetic COVID-19 data, we aim to build a visualization dashboard that features a choropleth map as well as a complementary chart.
- With the dashboard, we aim to allow users (both those experienced and inexperienced with public health methods) to explore the trends and geographic spread of COVID-19 in Massachusetts.
- All users must be able to operate a web browser, and intuit dashboard functionalities based on the dashboard labels (i.e. they need to be able to understand that they should click through a dashboard framework).
- Users can make educated and rational conclusions using the dynamic, interactive visualizations.
- More experienced users or those who are more curious would be able to explore the data with finer detail by extracting our specific features of interest, e.g. death counts, rates, hospital locations.
- We do not aim to implement an upload feature allowing users to import their own COVID-19 data. Instead, we seek to create an open exploratory visualization tool framework that allows users who have basic Python, GeoJSON and data cleaning knowledge to make visualizations from their own COVID-19 demographic data in the Observational Medical Outcomes Partnership (OMOP) common data format, and include an appropriate GeoJSON, with minimal effort.

Data:
Our dataset is a synthetic COVID-19 created by Synthea that was later converted into the OMOP common data model. The data can be found at:
https://forums.ohdsi.org/t/synthetic-data-with-simulated-covid-outbreak/10256

The data spans a period of three months, January 2020 to March 2020, mimicking the start of the pandemic and contains approximately 10,000 unique patients. The OMOP model is used and thus we have the following data tables at our disposal: `cdm_source`, `condition_era`, `condition_occurrence`, `death`, `drug_era`, `drug_exposure`, `location`, `measurement`, `observation_period`, `observation`, `person`, `procedure_occurrence`, `visit_occurrence`.

Each of the data tables have their own keys, but can be joined by `person_id`. However, we do not plan to utilize all of the available data tables since many are irrelevant to our use cases. Our analysis will be within the `condition_occurrence`, `death`, `location`, and `person` `tables`. Within the `person` table, we will be able to obtain the gender, race, ethnicity, and death date (if applicable). The `location` table contains the physical address of the patient. However, although there is a ZIP code field within the `location` table, we have estimated that about 50% of the column is not available; therefore, we will be utilizing the county field instead. COVID-19 information is stored in the `condition` table. Using the Athena vocabulary standard, we have determined that a `condition_concept_id` of '37311061' indicates 'Disease caused by 2019-nCoV' (the virological name of the virus that causes COVID-19).

Additionally, we will also be incorporating the latest United States 2010 Census data and the location points (latitude and longitude) of hospitals in Massachusetts. To fully visualize the data, we will be using a GeoJSON file of Massachusetts representing county boundaries. Ideally, geographic-demographic data table joining is completed based on Federal Information Processing Standard (FIPS) code. However, our synthetic dataset does not include FIPS codes, and so we have opted to perform GIS joins by simple county name (i.e. "Hampshire", not "Hampshire County"). The particular GeoJSON file we are using is found at the following Github Repository of topographic GeoJSON files for open use:
https://github.com/deldersveld/topojson/tree/master/countries/us-states

Users:
We are target health analytics personnel interested in visualizing the OMOP data format synthetic COVID-19 data for the state of Massachusetts. Currently, visualizing and understanding COVID-19 data is a priority for many in the healthcare field. However, technical skills within this field are variable. We believe that health analytics personnel or any other interested healthcare professional with limited to no experience with Geographic Information Science (GIS) or technical skill would be able to benefit and take advantage of the dashboard. No coding or querying would be necessary allowing a larger group of users to take advantage of the dashboard. Additionally, these users may be inclined to present the data to a larger audience of varying levels of expertise. Using this dashboard would allow them to provide and

convey the numerical data visually and translate the data in an understandable and intuitive manner.

- *Use case #1:*
  A health analytics personnel is interested in exploring the density of COVID-19 cases on the county level for Massachusetts. The user in this use case may have a goal of understanding trends and patterns of the outbreak or may simply want to explore the data in a visual manner. Additionally, the user would not necessarily need complex technical skills as the visualization would default to displaying a basic aggregate choropleth and corresponding line chart or histogram of overall case counts. The user would simply open the dashboard and interact with the visualizations; no uploading of data would be necessary, increasing the ease of use and accessibility of the tool. The user can further explore the data through the interactive features that are present (e.g. zoom in and out, export the map in an image format, hover over counties to get more specific county-level information). Once the user is finished using the tool, he or she can simply access the dashboard again for future use, or export the visualizations to be used in presentations or publications.

- *Use case #2*:
  A health analytics personnel or public health enthusiast interested in visualizing their own COVID-19 dataset, whether it be more rich and detailed synthetic data, real COVID-19 data, or data from a different state. This user would be interested in finding and utilizing a “plug-and-play” solution to visualize data in a manner that would require minimal effort. The user would need to download or clone or repository containing our step-by-step README file. The repository would include a set of GeoJSON files allowing the user to use a different state, if necessary. Otherwise the user can edit the lines importing our synthetic Massachusetts data and replace it with the paths to their own dataset of interest. If the user’s dataset is not properly formatted, he or she could use our SQL query and python scripts to format it appropriately from the OMOP common data model. Once the scripts import the desired data, the user can perform similar exploratory tasks as use case #1.

- *Use case #3:*
  A curious and aspiring data scientist interested in learning Plotly and Dash. This user may be interested in creating their own visualizations for data, particularly COVID-19 data, and may not have the resources or examples to pursue the task. This user may use our visualization tool as a stepping stone and example to learn how to create choropleths and epidemiological graphs using a common data model frequently used for COVID-19 data.