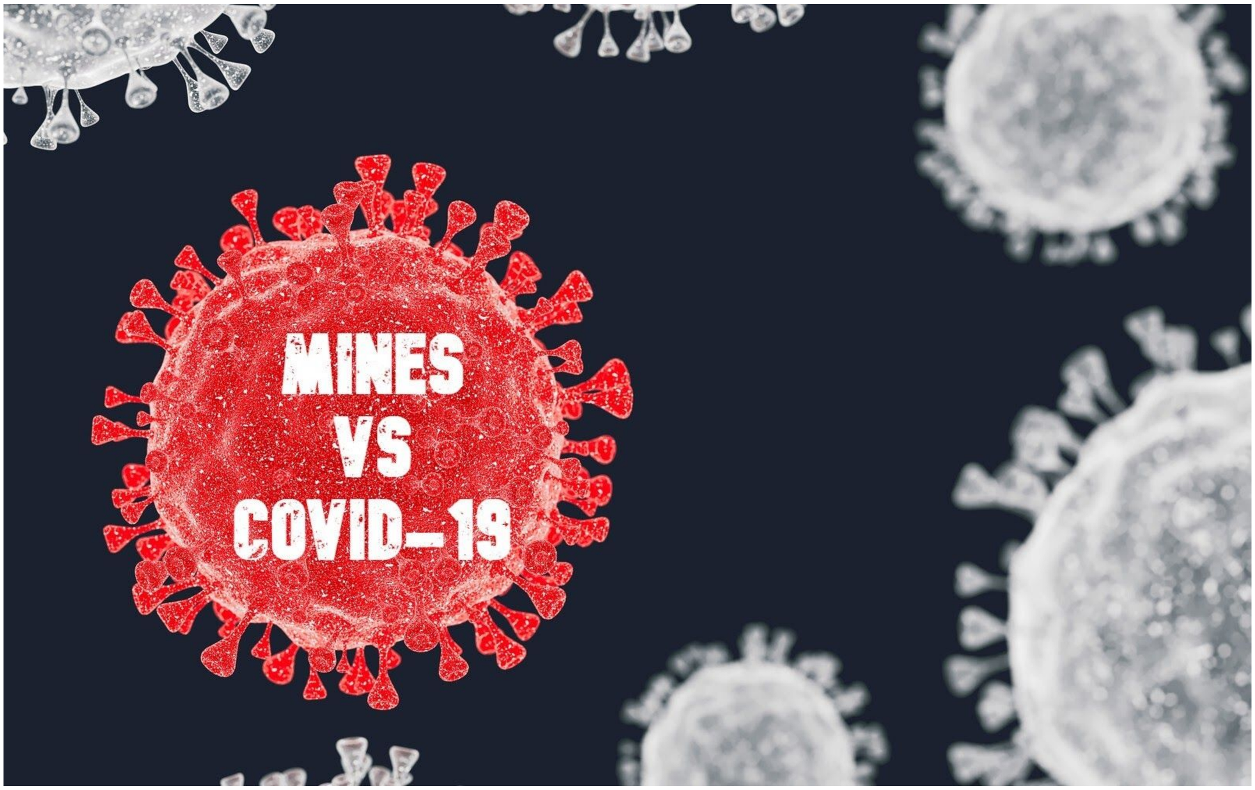
# Prediction of Coronavirus Infection and Complications in Individuals

Final Report

Class: CSCI 370 - Team: Klein 1 - Advisor: CPW
Nelson Hogan - Grace Petryk - Daniel Vigil - Shelby Cornelius



## Introduction

The coronavirus (COVID-19) pandemic has become a worldwide crisis, and novel tools are needed to reduce the healthcare and social burdens of this disease. Despite precautions such as staying at home, wearing masks, and social distancing, COVID-19 continues to spread. In many cases the disease causes severe complications, including death. Our initial project goal was to create two predictive models that would be the driving force behind a user-friendly COVID-19 app. The first model would predict COVID-19 infection (yes/no) in an individual, based on user input regarding symptoms, lifestyle factors, demographic factors and other health conditions. The second model would predict whether a COVID-19 infection would be mild or severe.

These models would be created by applying machine learning methods to individual-level healthcare and proteomic data. They require training on data from multiple sources related to COVID-19 cases. These sources include healthcare databases that have data corresponding to: patient infections of COVID-19, lifestyle factors, demographic factors, and other health conditions. Examples of lifestyle factors include diet, salt intake and smoking. Examples of demographic factors are age, gender and ethnicity. Other health conditions include high blood pressure, diabetes, chronic obstructive pulmonary disease, coronary heart disease, cancer and chronic renal disease. Sources also include genetic/proteomic databases that have information about genes and proteins related to COVID-19 and other health conditions. For example, the Online Mendelian Inheritance in Man (OMIM)[1] database has compiled a list of human proteins that are targeted by COVID-19, and the DisGeNET[2] database has data on associations between certain proteins and health conditions. Thus, a major key to developing the model is gathering and processing the large amounts of data specified by our client. This includes applying for data, retrieving data, and managing the data in a semi-permanent mobile database. All of this training culminates in a model that can accurately predict COVID-19 infection and complication likelihood based on new user-submitted inputs.

The COVID-19 application would require user input regarding the user's personal health factors identified as important for predicting COVID-19 infection and/or severity of infection. This input from the user would be processed by the models to generate the likelihood of infection and complications, and may include advice for the user based on predictions. The intended audience is the general public, healthcare professionals and other researchers. This would likely require two different versions of the app, but both would use the predictive models. Our primary goals were obtaining the data and developing the predictive models. However, the time required by the initial steps of the project proved to be too long and the predictive models proof of concept and developed models were never built by this team. Our client will either use a professional app development company or a future team to build the models and create a GUI that can be deployed and available to the public.

While many COVID-19-related products and tools have been created, ours will venture into unexplored-powerful territory. Artificial intelligence/machine learning (AI/ML) has been scarcely used in the inflooding of Coronavirus related apps. In addition, the use of AI/ML trained on lifestyle, demographic, health history and genetic/proteomic data has not been done before and can provide helpful predictions with potentially higher accuracy than other apps on the market.

## Requirements

### Functional Requirements

- Submit data requests to obtain access to non-public datasets
- Integrate and analyze data from multiple data sources:
  - Molecular data on coronavirus protein targets and genes/proteins associated with other medical conditions to identify comorbidities that may be used as predictors
  - Clinical data to identify possible predictors of coronavirus infection and complications, among demographic, medical history, exam and laboratory data
  - Data obtained from prior research
- Data-trained mathematical model
  - Takes input regarding an individual such as demographic, symptoms, comorbidities, etc.
  - Produces an output giving likelihood of contracting COVID-19 and of developing complications

### Non-Functional Requirements

- Visualizations and conclusions from any model are easy to read and understand
- Aggregated data is easy to work with
  - Uses SQL
  - Available to other researchers and future teams (as determined by our client)
- Mines provided PostgreSQL server
  - Transferrable
- No costs expended outside of grants

## Architecture

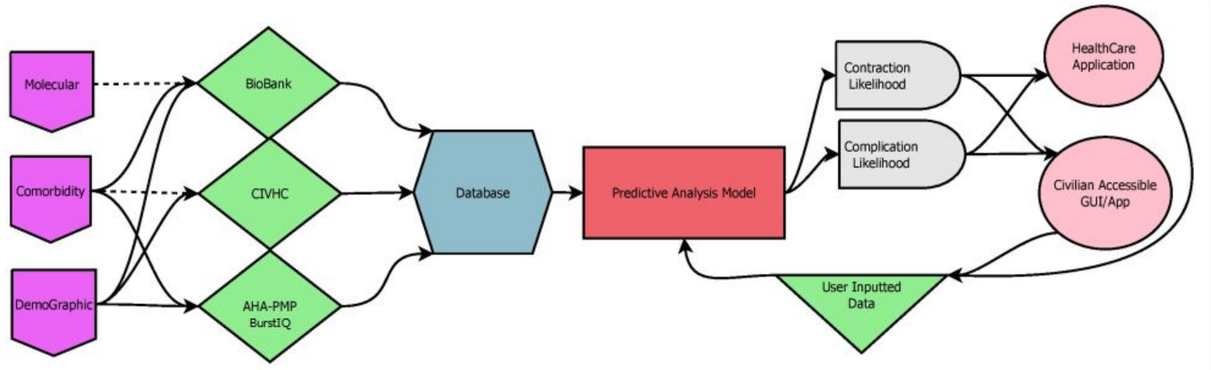


Figure 1: Architecture Diagram

While the timeline and expectations of the project changed dramatically, the overall architecture created by the team did not. The data sources shifted to be the main goal of the team to enable the right half of the architecture diagram seen in Figure 1.

As seen in Figure 1, the sources of data (green) each correlate to different types of information gained (purple). These are concatenated into the database(blue) as a central hub of data for the use in training and developing the predictive analysis model. Branching off the model, the outputs (grey) feed into the intended users (pink) whose responses can feed back into the model, training it further (green triangle).

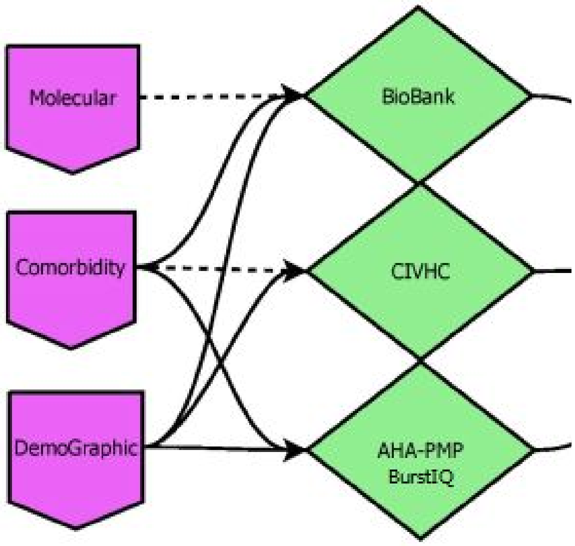


Figure 1 (detail)

This foundational architecture is a high-level view of what is to be accomplished, with the team only having time to complete some of the left-hand side with database creation and data access.

Honing in on one component and its relation to the data, BioBank[3] is a prime example. Looking for its connection points to the left, BioBank contains Demographic, Comorbidities (other health conditions), and Molecular data. UK BioBank is a comprehensive study of around 500,000 people from the UK and their medical conditions, biomarkers, health, etc. over the span of their life. The solid lines extending from Demographic and Comorbidity indicate that the data is accessible and processible. The dashed line extending from the Molecular block means that the data is there, but currently hard to access or use. In the case of BioBank, the molecular genome data was formatted in such a way that it was not viable for the team to use, however, it is included in the architecture as this is a precursor project and the next team to pick up the project could elect to use the complex data, should they want to devote more time and resources there.

Once we have access to the data, the next step is to store it in a PostgreSQL database and do some data science on it. As mentioned before, this server is provided by CCIT. The server currently runs CentOS, has 4GB of RAM, 30GB of storage, and 2 processor cores with a base frequency of 2.1 GHz and can turbo up to 3.7GHz. The server as it is now is underpowered for data science work, since when it was requested it was only intended to be used for data storage. Fortunately since the server is virtualized it is possible to expand this later without much need for migrating what's already on the server. For manipulating and understanding the data we're using Jupyter notebooks hosted on the server. We chose this because it allows us to rapidly see results while we're working with the data, as well as providing us with a shared space to work.

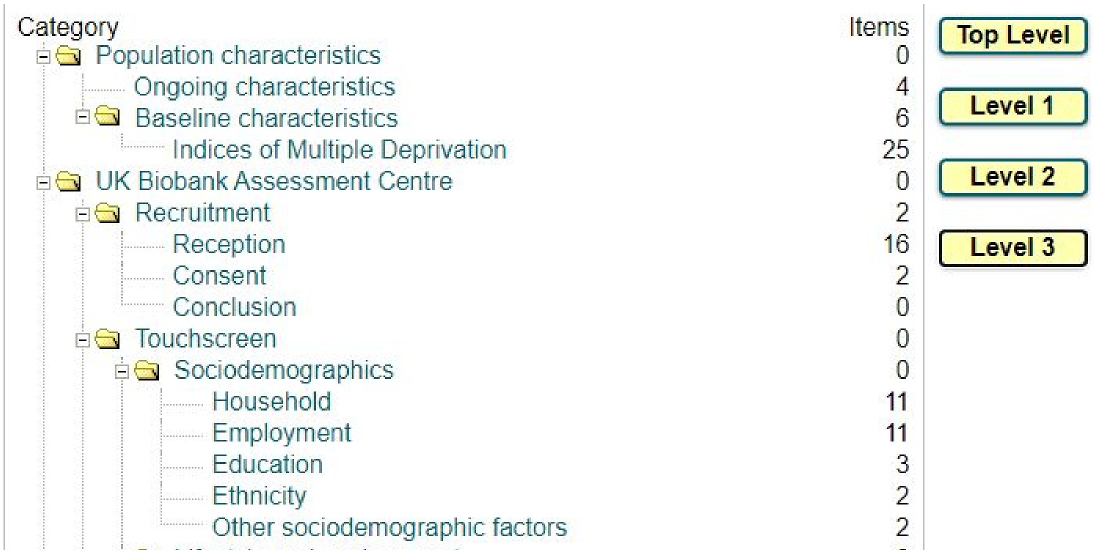
By unifying all the data we collect in a single place, we're making it far easier for whoever works on the project next. By far the most time-consuming part of the project has been aggregating and collecting all the data from varied sources.  By centralizing the data with a simple and consistent database structure, the next team to work on this project will not need to do the tedious work of collecting the data in the first place.

# Technical Design

## UK BioBank

One of the most unique data sources used was the UK BioBank. The BioBank is a comprehensive database that has tracked and is currently tracking the medical status of approximately 500,000 UK citizens. This began decades ago and has kept up with medical changes of the participants while also containing genomic sequencing of the participants. Recently, the results of approximately 1,500 COVID-19 tests conducted on the BioBank participants have been entered into the database.

When interacting with the BioBank, an application must be filled out detailing why the data is being requested and specifying the data requested. The data request is held in a data basket that is handpicked by the researchers from the data in their showcase. This showcase is broken down hierarchically, as seen in the figure below.



| Category | Items |
|---|---|
| Population characteristics | 0 |
| Ongoing characteristics | 4 |
| Baseline characteristics | 6 |
| Indices of Multiple Deprivation | 25 |
| UK Biobank Assessment Centre | 0 |
| Recruitment | 2 |
| Reception | 16 |
| Consent | 2 |
| Conclusion | 0 |
| Touchscreen | 0 |
| Sociodemographics | 0 |
| Household | 11 |
| Employment | 11 |
| Education | 3 |
| Ethnicity | 2 |
| Other sociodemographic factors | 2 |

After the desired data fields are selected, filters for said data are very much encouraged. In this project's case, filters were selected that limited the data rows to only participants that:

1. Had been tested for COVID-19 (positive or negative result)
2. Had all vital measurements (none were null)
3. Certain comorbidity measurements were not missing (such as asthma or smoking history)

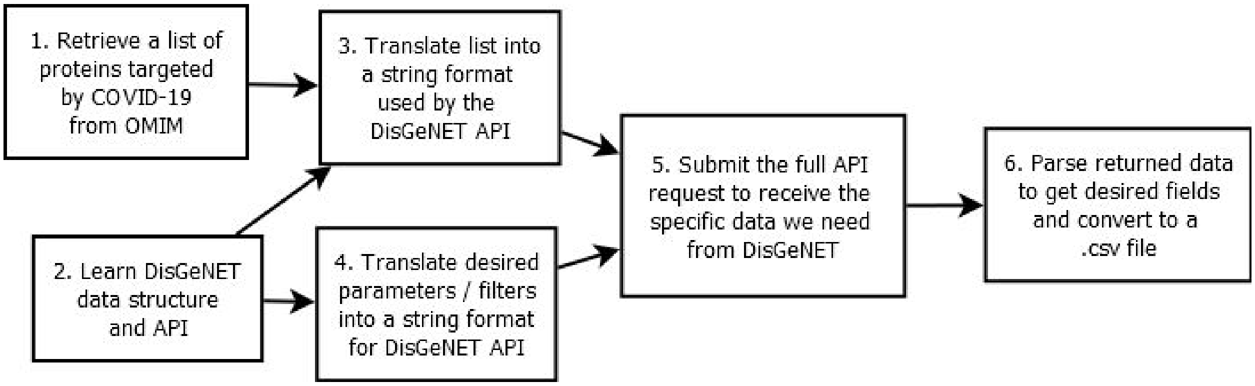
These general filters provided for approximately 1,520 rows of participants' data and between 200-400 columns of data (the application was approved but the payment is yet to be processed so the numbers are not concrete).

The downside to UK BioBank data is that the scope of this project is focused around America as far as demographics go. The location selected by the user will likely be state/county and if a person is located outside of America (all of the BioBank participants) then the model will produce a less accurate prediction; this is due to the proximity and density of COVID-19 being so influential in risk. An additional downside, it is unlikely there will be many participants in the BioBank who tested positive for COVID-19 when considered statistically.

**DisGeNET Data Retrieval**

One important tool for data retrieval is Application Programming Interface (API). An HTTP based API can be used to request and receive specific data from a source using a detailed URL. Traditionally, sources have provided data in pre-created documents or files. But an API can allow the user to apply filters, set parameters or use SQL-type queries to retrieve the data they need. We used APIs for two of our data sources: DisGeNET for associations between proteins and health conditions, and BurstIQ [4] for COVID-19 related data. Here we will describe our process around retrieving and using the DisGeNET data.

We used the DisGeNET data by cross-referencing it with data from the Online Mendelian Inheritance in Man (OMIM) database about proteins in the human body that are targeted by COVID-19. This ultimately generated a list of health conditions that are associated with the proteins targeted by COVID-19. We completed most of this using Python and relevant modules. The key steps are shown in this flowchart:


1. Retrieve a list of proteins targeted by COVID-19 from OMIM
2. Learn DisGeNET data structure and API
3. Translate list into a string format used by the DisGeNET API
4. Translate desired parameters / filters into a string format for DisGeNET API
5. Submit the full API request to receive the specific data we need from DisGeNET
6. Parse returned data to get desired fields and convert to a .csv file

Step 1: Our client provided an Excel file she had previously obtained from OMIM containing proteins targeted by COVID-19 and related data. We isolated the proteins, converted it to a csv file, read it using Python's 'csv' module and eliminated duplicates.

Step 2: We learned DisGeNET's data structure and the meaning of different data fields from documentation on their website. We learned their API format using their Swagger page. Swagger is a documentation tool for HTTP based APIs. It allows you to enter and submit parameters, then it generates a URL fitting that site's API and returns the results. By entering different parameters and studying the pattern of these URLs, we learned their API format.

Step 3 & 4: Using the patterns we derived from DisGeNET's Swagger page, we iterated through the set of proteins to create a concatenated string, and converted parameters to the proper format. These were combined into URLs fitting DisGeNET's API, which were used in step 5. This allowed us to designate which proteins we were interested in and minimum thresholds for strength of evidence.

Step 5: We used Python's 'requests' module, which is designed for making API requests via code, to retrieve data from DisGeNET. These requests used the URLs created in Steps 3 & 4 and returned records meeting the designated criteria. The results were put in JSON format using Python's 'JSON' module, to make them easier to work with.

Step 6: We put the records into a dataframe using Python's 'pandas' module and iterating through the JSONs created in Step 5 to extract the fields we needed. Finally, we exported the dataframe as a csv file to use in the next segment of the project.

## Ethics and Quality Assurance

This section contains first a list of relevant ACM/IEEE principles that were maintained during the project, or must be observed during the future stages of the project when picked up. Below those principles is an indexed list of practices employed during the project [Ethical Tree]; in reference to this list the Task-Ethics Correlation table describes tasks achieved or to-be-achieved and the index of ethical practices which apply.

**Most Pertinent ACM/IEEE Principles**

- 2.2 Maintain high standards of professional competence, conduct, and ethical practice.
- 3.1 Ensure that the public good is the central concern during all professional computing work.
- 3.7 Recognize and take special care of systems that become integrated into the infrastructure of society.
- 3.13. Be careful to use only accurate data derived by ethical and lawful means, and use it only in ways properly authorized.

The above ACM/IEEE principles are important due to the current state of the world due to the COVID-19 pandemic. This tool is meant only to benefit the public. By ensuring that data is obtained ethically, our tool follows HIPAA guidelines as well as reassures the public of the trustworthiness of the application they are using. Also, by recognizing the integration of this tool into the public, that requires the public good to be our top priority. This kept our group from creating false data for a proof of concept as that proof of concept may be introduced to the public under the guise of true data, once the project is out of our hands.

**ACM/IEEE Principles Most in Danger of Violation**

- 1.04. Disclose to appropriate persons or authorities any actual or potential danger to the user, the public, or the environment, that they reasonably believe to be associated with software or related documents.
- 1.06. Be fair and avoid deception in all statements, particularly public ones, concerning software or related documents, methods and tools.
- 1.07. Consider issues of physical disabilities, allocation of resources, economic disadvantage and other factors that can diminish access to the benefits of software.

These ACM/IEEE principles are in danger of violation without vigilance on our end as software engineers. The world is in peril to an extent which we have not seen in our lifetime. Our tool must ensure ease of access to people of all income levels and demographics. These principles are most in danger as we are handing off our project to others, as they may not prioritize our ethical vision. However, the work done within the scope of this project will uphold these important ethical values outlined.

**Michael Davis Tests**

Harm test: Does this option do less harm than any alternative? Do the benefits outweigh the harms?

> Our project may incorrectly determine treatment for COVID-19 due to the yes/no nature of the predictive tool. The tool may suggest that you are not at risk for severe complications even though you are very close to the cutoff which does not negate the fact that you could have complications, but may create a false sense of security and have negative implications.

Reversibility test: Would this choice still look good if I traded places? (i.e., if I were one of those adversely affected by it?)

> We may be upset that we put our trust into a tool that provided us with a false sense of security, we would also be aware that the tool cannot replace the medical advice of a medical professional, so we may take recommendations with a grain of salt and at the end of the day consult with healthcare professionals.

**[Ethical Tree] Ethical Considerations for Quality Assurance**

1. Data Ethics
    a. No illegal gathering of data
        i. Proper data access procedures are used for each data source
    b. Data integrity
        i. Rights to acquired data stay with Judith Klein (Client)
        ii. Data access is limited to the team and the client
        iii. Security of data is ensured
            1. Centralized database for housing data
            2. No local permanent storage of data
    c. HIPAA Compliant

        i. Each data source maintains HIPAA compliance - this extends to data received from the source so database is HIPAA-compliant

2. Project Timeline Assurance
    a. Client is kept in the loop with expected timelines
        i. Timeline on this project has changed dramatically -- all in line with client expectations and understanding
    b. Client is included in our project Slack channel and is kept updated on progress and challenges
3. Potential Harm of Project
    a. Should model be only initially developed
        i. Will not be approved without proper testing
        ii. Will be set up for testing and further development
    b. Improperly developed model
        i. Can result in user not seeking help for potentially deadly Coronavirus
        ii. Can result in user seeking unnecessary help for Coronavirus
            1. Hinders hospital effectiveness
            2. Possible negative financial and emotional impact on the user
4. Code Quality
    a. Pair work has been used at multiple stages
        i. Data acquisition
        ii. API usage
        iii. Database creation
    b. Used proper resources
        i. Team integrated into BurstIQ development team slack channel because the BurstIQ database is complex and was not understood well
    c. Accurate model
        i. Statistical analysis of results must meet a certain standard
            1. If the standard is not met the model cannot and will not be released. It will be held as a reference point for refining and a proof of concept
    d. Code is open for development but closed for modification
        i. The model when complete can be implemented in an app
        ii. Model will be modifiable until it is statistically acceptable
5. Data Quality
    a. Data will not be used if
        i. Obtained unethically
        ii. Source is not reputable
        iii. Not understood by team
        iv. Not applicable

**Task-Ethics Correlation Table**

Achievable/Achieved-Not achievable in current project scope

| Task | Effect | Reference to Ethical Tree section above |
|---|---|---|
| Data Collection | Multiple data source requests and acquisition work emplaced | 1.a<br>1.b.i<br>2.a.i<br>5.a.i<br>5.a.ii |
| Database Management | Comprehensive, intuitive, and secure database is created and passed along | 1.b.ii<br>1.b.iii<br>4.a.iii<br>4.d.iii |
| Initial/Surrogate Model Creation | Create a "starter" model that can mimic the actual model to accelerate and debug predictive analysis on COVID-19 data | 1.a<br>1.b.iii.<br>2<br>4.a.iv<br>4.c<br>5.a.i<br>5.a.iv |
| Data Features Identified for Use in Model | *Expected* data features and categories pertaining to statistical model have been identified for use in creating model | 5.a.ii<br>5.a.iii<br>4.c<br>3.a |
| Statistical Model Creation | Model based on acquired COVID-19 data has been created and is ready for testing | 2.a<br>3<br>4.a.iii<br>4.c<br>4.d.i<br>4.d.ii |
| Populate Database with COVID-19 Data | Multiple sources of COVID-19 data are processed and database is filled with the meaningful columns and rows | 1.b.iii<br>1.c<br>4.a.i<br>4.a.iii<br>5.a |

## Results

The initial high-level goal of this project was to implement a machine learning tool derived from multiple data sources related to COVID-19 cases and its comorbidities. This tool would be used to create binary predictions based on user input for: comorbidities, lifestyle factors, and demographics. This input would allow for prediction of a binary outcome (yes or no) of COVID-19 infection, and prediction of complication likelihood (serious or mild symptoms). Unfortunately, data acquisition has been a long and tedious process that blocked our team from being able to implement a machine learning tool with all of the data sources preferred by the client. Overcoming the bureaucratic struggles, the team was successful in gathering and maintaining data from multiple sources, and developing specific processes and plans to facilitate the continuation of the project. This culminated in data sets from three different sources, with more pending, and the establishment of a database that will be passed onto the next group working. The work completed presents a solid foundation of data, structures, and ideas/plans to be used by a future team. The duration of data acquisition which prevented the team from doing any predictive analysis on the data was of unforeseen and unfortunate length. This was acknowledged by the client as unavoidable.

The future of this project will stand upon the foundation laid here. While the initial goals of the project became unachievable, the project instead became the foundational work that will enable the next team to develop the model and app without issue.

The learning in this project was derived from high-level real-world problems (as opposed to low-level coding problems). An example of real-world learning was that having data already provided is essential to meaningful data analysis and manipulation; while that statement seems obvious, it is more complex than meets the eye. Without immediately accessible data, data acquisition is long, costly, and often requires learning new techniques and processes. Throughout the process, however, the team learned concrete foundations for high-level data science. Each data source was diverse in its data acquisition and problems it presented: BioBank required applications and data basket gathering including data selection and filtering, BurstIQ had their own SQL variant called TQL that members of the team had to learn, AHA-PMP [5] had data presented in API’s which several members learned how to work with, and DisGeNET required members to learn how to work with proteomic data.

Overall, the work done in this field session project required the team to be dynamic and persistent. The team learned along the way and created solid legs for this project to run on. The full potential of the project is yet to be fulfilled but will be with the work done here.

## External Organization References

[1] Online Mendelian Inheritance in Man (OMIM), *National Center for Biotechnology Information, U.S. National Library of Medicine*
https://www.ncbi.nlm.nih.gov/omim
[2] DisGeNET, 2010-2020
https://www.disgenet.org/dbinfo
[3] UK BioBank, 2019
https://www.ukbiobank.ac.uk/about-biobank-uk/
[4] BurstIQ, 2015-2020
https://www.burstiq.com/company/
[5] AMA-PMP, *American Heart Association*
https://precision.heart.org/about