Project Deliverable 2

Pandemic Insights: A Comprehensive COVID-19 Retrospective Analysis Tool

Group 7

Members

Vivek Milind Aher	vaher@ufl.edu
Dhruv R Makwana	dmakwana@ufl.edu
Sakshi Pandey	sakshi.pandey@ufl.edu
Vishal Prakash	vishalprakash@ufl.edu

Overview and Description of the Application

The COVID-19 pandemic, a crisis of unprecedented magnitude, has left an indelible mark on humanity. Its profound impact on various aspects of human life necessitates a thorough retrospective analysis to uncover the factors and circumstances that shaped its trajectory. This application has been conceived to empower a diverse range of stakeholders, including healthcare professionals, policymakers, environmental scientists, and researchers, with the tools and insights required to conduct comprehensive evaluations of the pandemic.

At its core, this COVID-19 application seeks to enable users to delve deep into the pandemic's multifaceted nature. It aims to provide valuable resources for those who can affect change and drive future policies based on the lessons learned from this global health crisis. Whether it's doctors aiming to understand comorbidities' impact on mortality rates, policymakers striving to make informed decisions, or researchers seeking to unravel complex pandemic-related trends, this application offers a versatile platform for in-depth analysis.

The application's functionality encompasses data computation, statistical testing, and dynamic visualization of interdisciplinary datasets. These features transform complex data into intuitive visual representations, allowing users to identify critical insights. With an interactive interface that enables the selection of parameters and time frames, users can explore diverse data trends, contributing to a more profound understanding of the pandemic's dynamics.

The database's structured design ensures efficient data management and accessibility. Drawing data from diverse sources and disciplines, it facilitates comprehensive cross-disciplinary analyses, providing holistic insights into the pandemic's various dimensions.

Complex trend queries play a pivotal role in this application, allowing users to uncover correlations and patterns related to the COVID-19 pandemic. These queries can illuminate the influence of socioeconomic factors and race on pandemic outcomes or shed light on shifts in air quality and their environmental implications. By harnessing the power of complex queries, this application can scrutinize resource allocation trends and pinpoint instances where the focus on COVID-19 may have inadvertently led to neglect in treating other health conditions.

In summary, this application's success hinges on robust database support and the capability to perform intricate trend queries. These features empower users to navigate the complex landscape of pandemic data, make informed decisions, and pave the way for innovative policies and practices, not only for future pandemics but also in the realms of environmental research and epidemiology. This application stands as a crucial tool in our ongoing battle against global health crises.

User Interface Design

The COVID-19 pandemic has presented unprecedented challenges to societies worldwide, demanding a concerted effort to understand its impact and chart a course for the future. In response to this global crisis, we have developed a comprehensive web application tailored to the unique needs of diverse user groups: the General Public, Epidemiologists, Doctors, Policymakers, and Environmental Scientists. Each section of this application serves a distinct purpose, equipping users with the tools and insights necessary to navigate the complexities of the pandemic.

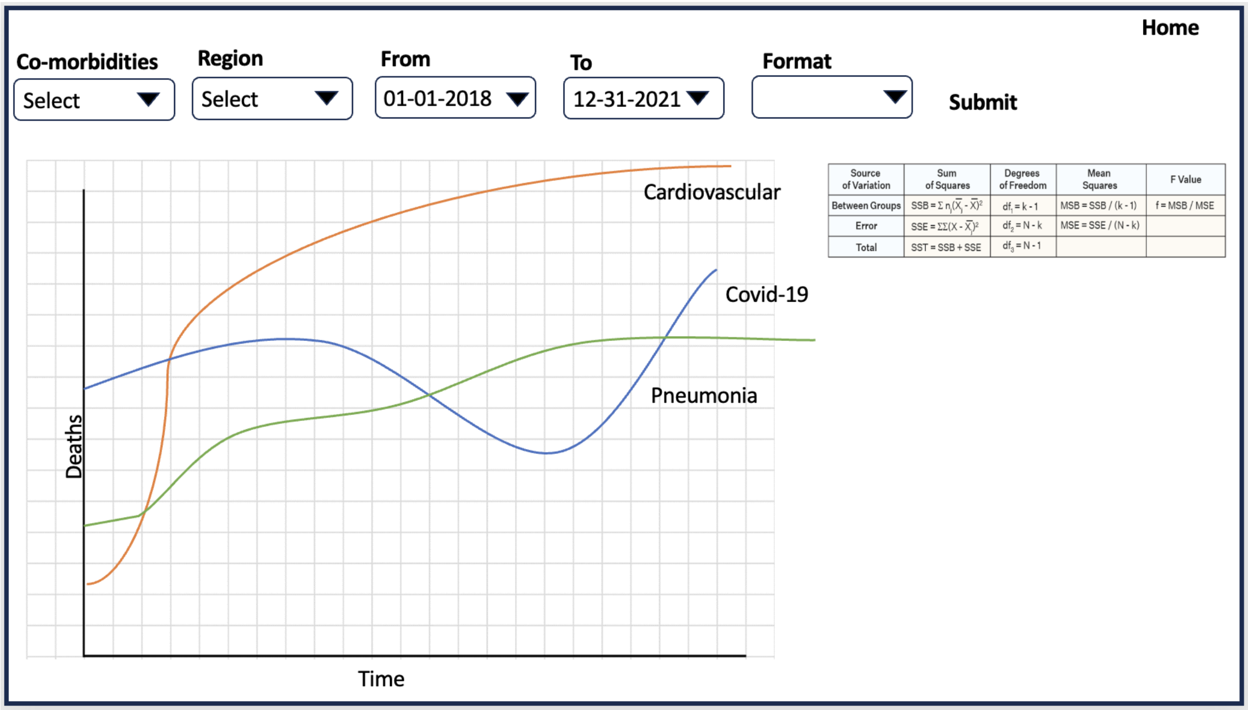
In essence, this web application serves as a central hub for understanding and navigating the COVID-19 pandemic's multifaceted challenges. It is our belief that by providing valuable data, analytical tools, and dynamic visualizations, we can collectively enhance our response to this crisis and lay the foundation for innovative strategies in healthcare, environmental research, and epidemiology. Together, we can shape a brighter and more resilient future in the face of global health crises.

Integration of the Complex Trend Queries

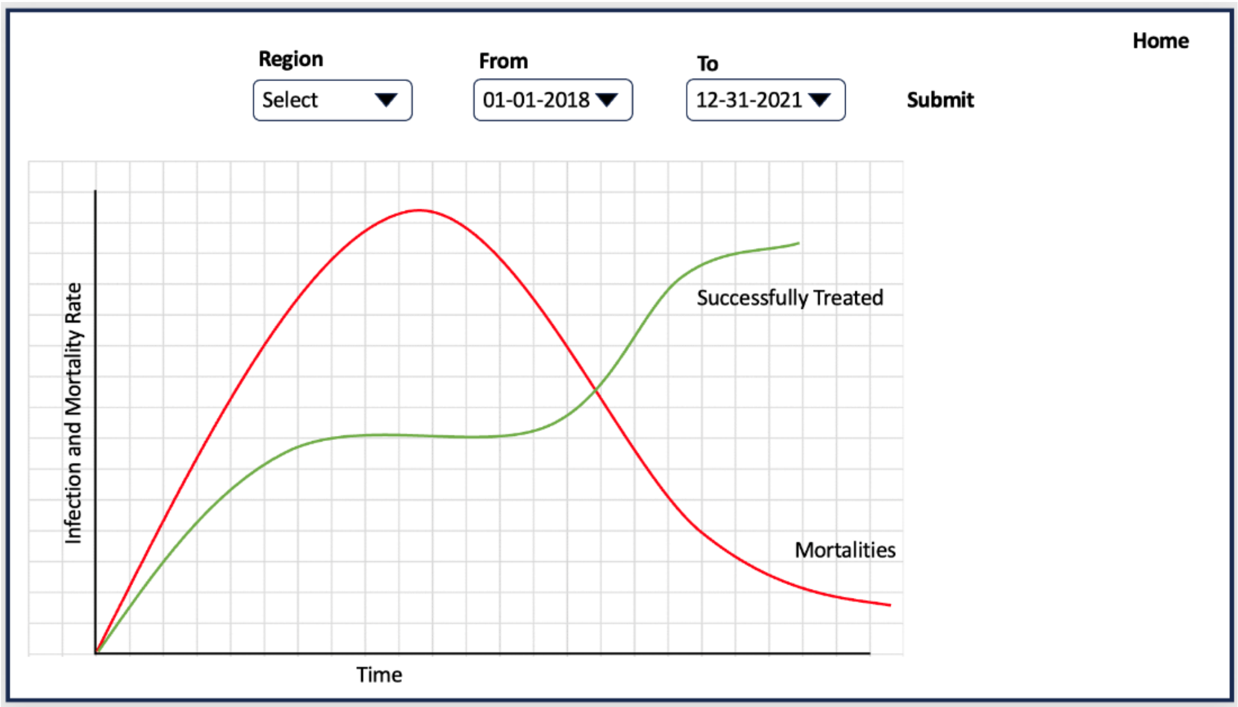
Query 1: How many people died due to other causes during the pandemic time period (cardiovascular death rate, other infectious disease death rate)? How many COVID-19 infections were successfully treated and how many COVID-19 patients died?

By comparing the means of successfully treated COVID-19 infections, COVID-19 mortalities, and deaths caused by cardiovascular reasons or car accidents using an ANOVA test for multiple samples, we aim to assess whether the allocation of hospital resources primarily towards the pandemic may have contributed to a higher overall mortality rate for other diseases.

This query can be used by multiple users including Doctors, Epidemiologists, and Policy Makers.



Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F Value
Between Groups	$SSB = \Sigma n_j(\overline{X}_j - \overline{X})^2$	$df_1 = k - 1$	$MSB = SSB / (k - 1)$	$f = MSB / MSE$
Error	$SSE = \Sigma\Sigma(X - \overline{X}_j)^2$	$df_2 = N - k$	$MSE = SSE / (N - k)$	
Total	$SST = SSB + SSE$	$df_3 = N - 1$		



The resultant wireframe therefore displays a visualization of the comparison between these factors, the table, and the results of the ANOVA test. An interpretation of the results detailing key insights and figures will also be provided to aid the user's studies and analyses.

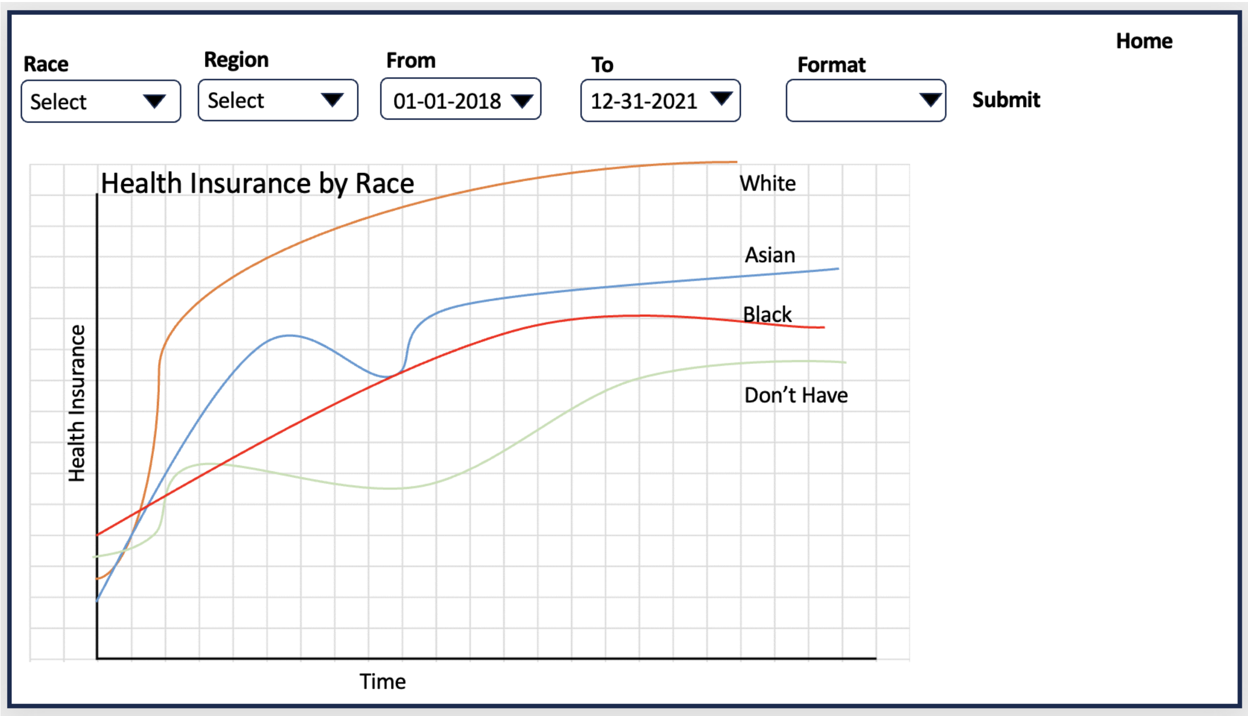
Query 2: What was the mortality rate of different races over time during the COVID-19 pandemic? For each region, which race had the highest mortality rate? How many people from that race had health insurance in each region vs. How many people from that race had no health insurance in each region?

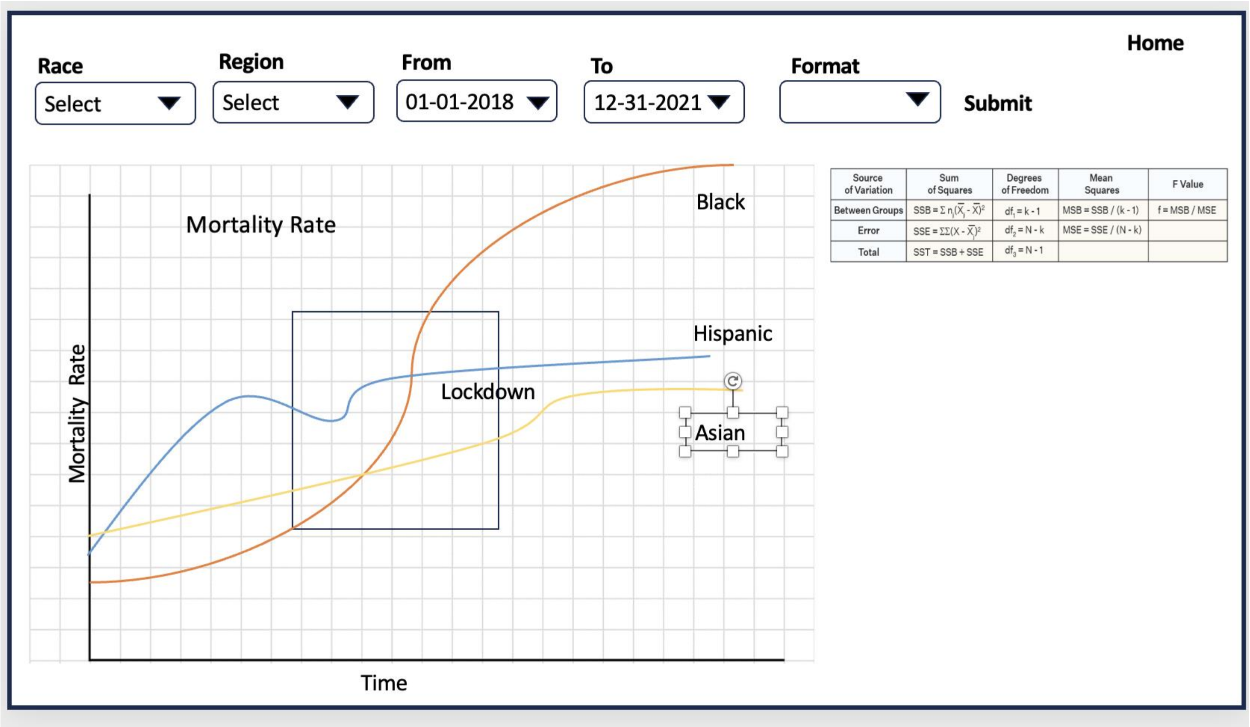
*Region (Northeast, Midwest, South, and West)

In this analysis, the user can investigate the mortality rates of different racial groups over a specific time frame during the COVID-19 pandemic. They can also identify which race had the highest mortality rate in each of the four regions (Northeast, Midwest, South, and West).

Additionally, it allows the user to delve into health insurance coverage within these regions. Specifically, the user can see the number of individuals from the highest mortality rate race who had health insurance and those who did not have health insurance in each region during the same time period.

The overarching goal of this query is to allow users such as Doctors, Epidemiologists, and Policy Makers to compare the racial group with the highest mortality rate to the racial group lacking health insurance coverage in each region. This analysis will shed light on the pandemic's impact on different racial communities and their access to healthcare resources across various regions in the United States.

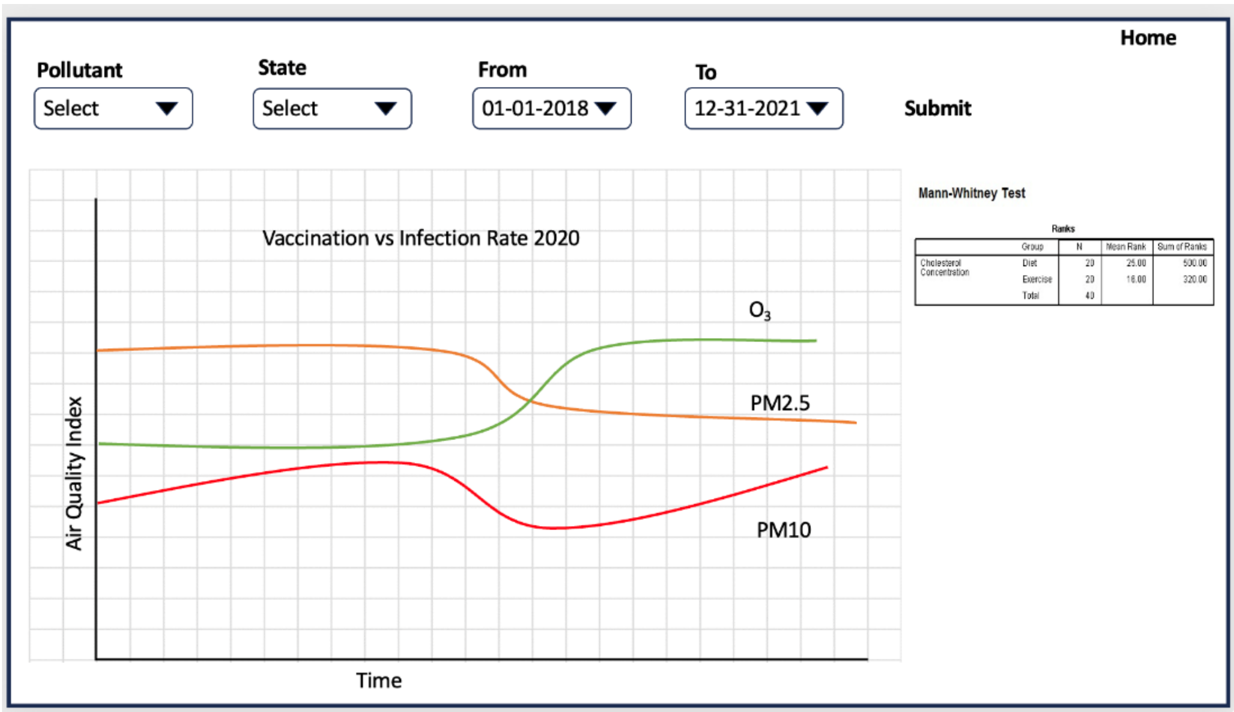




The resultant wireframe for this page is designed to display a comparative analysis between health insurance coverage across the nation and the mortality rates for different races in specific regions of the United States.

Query 3: Find the mean AQI for each pollutant over a period of time for each state. Following this compare the AQI data points (for a year prior to COVID-19 with a year during the pandemic, ex: 2018 vs. 2020) with the Mann-Whitney U-test to determine whether the points are significantly different or not?

This query provides a valuable tool for assessing the impact of the COVID-19 pandemic on pollution levels. By analyzing the mean monthly Air Quality Index (AQI) for various pollutants over a two-year period, users can gain insights into pollution rates before, during, and after the pandemic.



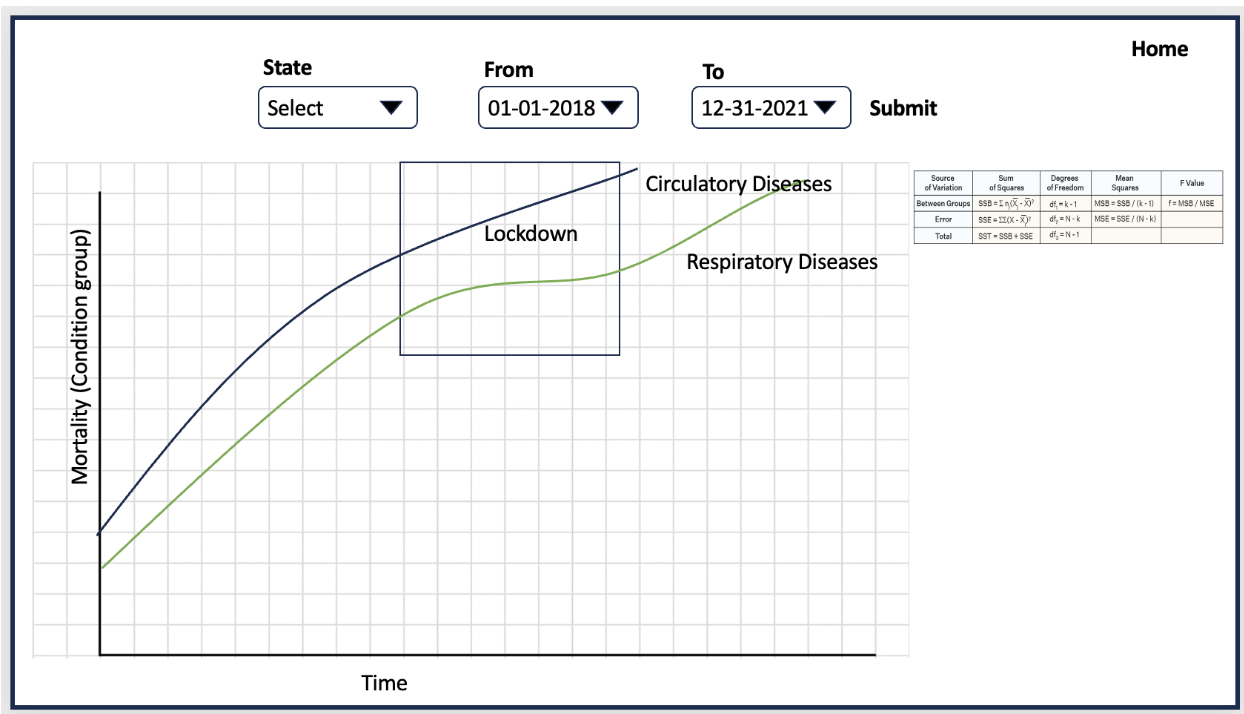
The resultant wireframe pictured above allows users to:

1. Select a specific pollutant of interest.
2. Choose a timeframe spanning two years, allowing you to compare the pre-pandemic, pandemic, and post-pandemic periods.
3. The query will calculate the mean monthly AQI for the selected pollutants within each of these timeframes.
4. Utilize the Mann-Whitney U-test to statistically compare these data points and determine whether pollution rates increased or decreased during the pandemic compared to the previous year.

Query 4: Over time which condition group (ex. circulatory diseases, respiratory diseases), in each state, led to how many mean mortalities per month? Compare these against one another using the One-Way ANOVA test and determine what is the difference in mortality with reference to the other condition afflicting the patients?

This query is a valuable resource for healthcare professionals and researchers seeking to understand the impact of COVID-19 on individuals with pre-existing medical conditions. It focuses on identifying which conditions or diseases were exacerbated following COVID-19 infection and ultimately led to mortality. Additionally, it provides insights into how specific diseases may contribute to fatality during a pandemic involving respiratory illnesses like COVID-19.

The interpretation of the results can be instrumental for healthcare professionals in understanding the risks associated with COVID-19 for patients with certain medical conditions. It can also guide researchers in healthcare to identify areas of concern during pandemics involving respiratory illnesses, aiding in the development of targeted strategies and interventions to improve patient outcomes.



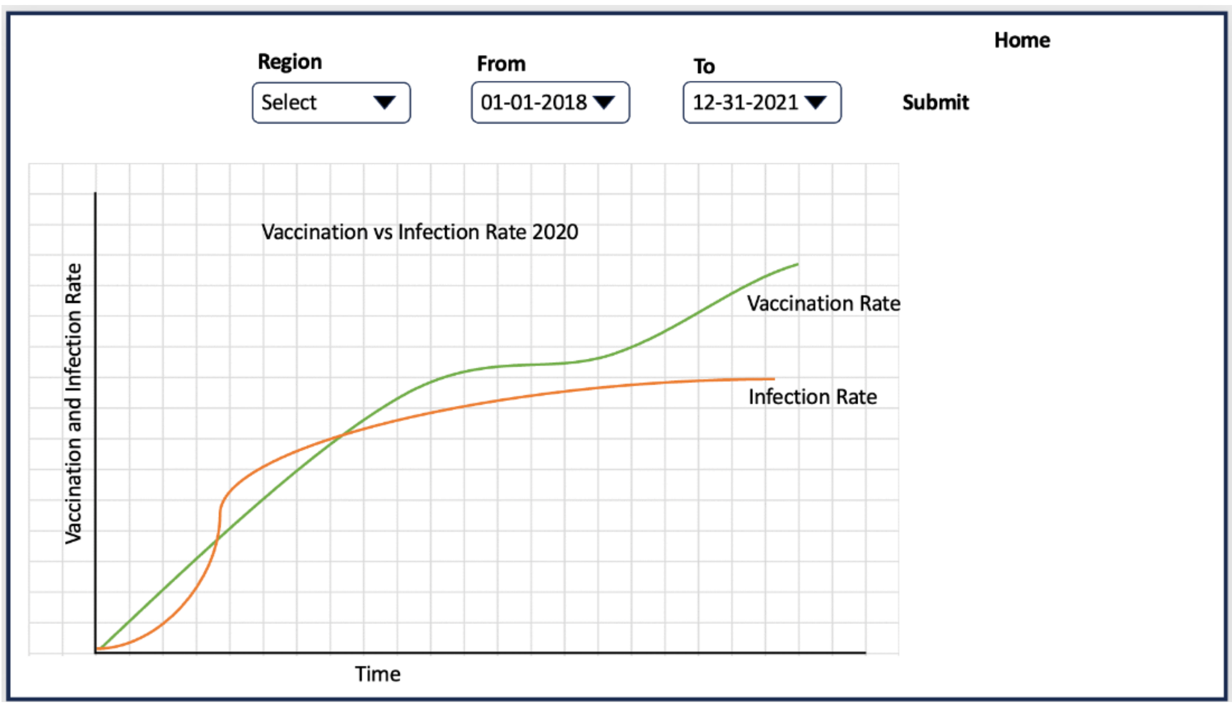
The resultant wireframe allows users to use this query to:

1. Specify the pre-existing medical conditions or diseases of interest.
2. Set a timeframe to analyze the impact of COVID-19 on individuals with these conditions.
3. Visualize the results and statistical tests.

Query 5: What was the vaccination rate vs. infection rate over time? How many people who were infected and tested positive had comorbidities?

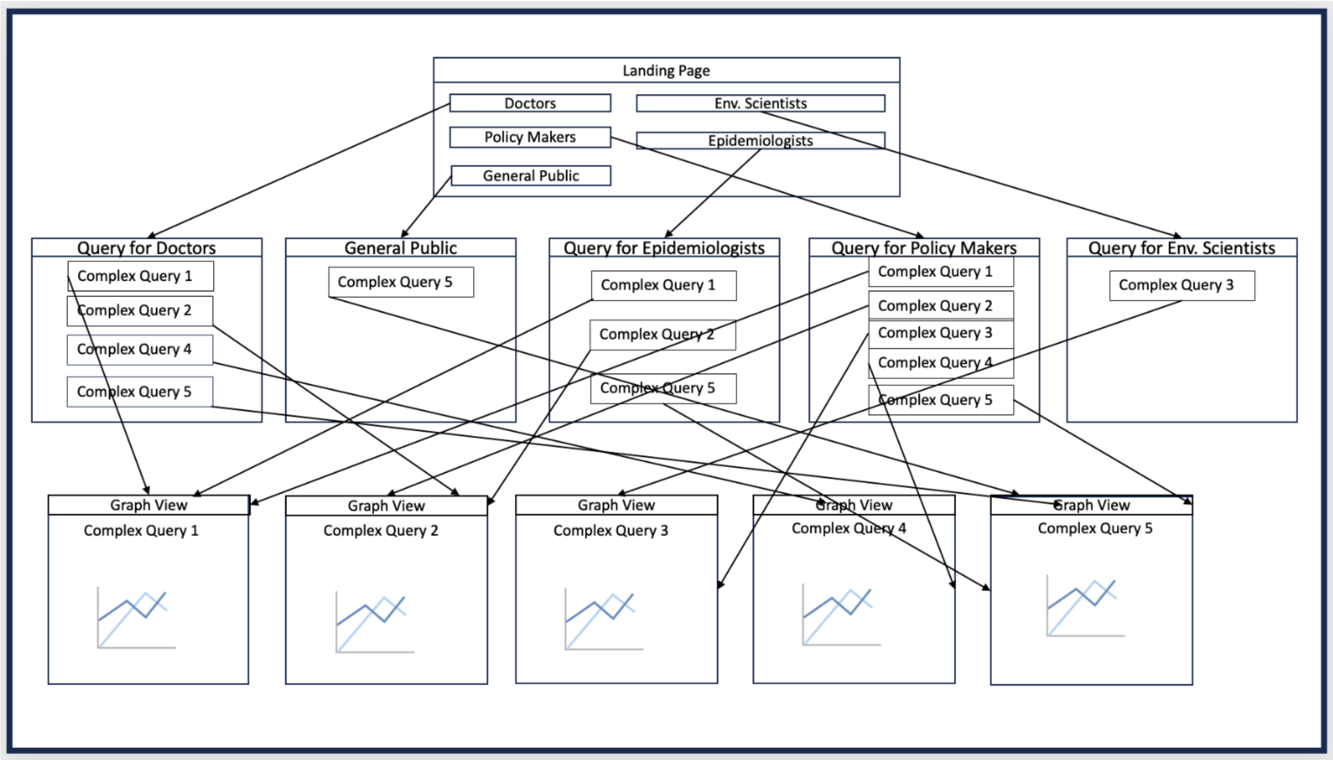
This query is mainly addressed to the General Public, Epidemiologists, and Policy Makers. It revolves around two core aspects of the COVID-19 pandemic: vaccination efficacy and the influence of comorbidities on infections and mortalities. The resulting insights from analyzing these datasets can aid researchers of various disciplines in gaining a more profound understanding of vaccination efficacy and its implications in the fight against the pandemic.

The goal of this query is to allow users to track the vaccination rate versus the infection rate over a specified period of time and delve into a critical examination of individuals who were vaccinated but later contracted COVID-19.

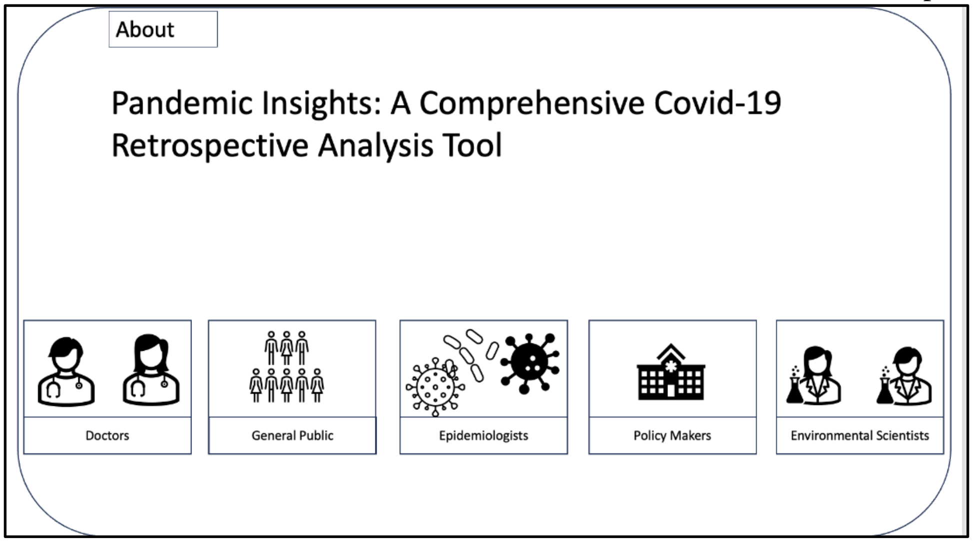


The resultant wireframe allows the users to scrutinize the presence of comorbidities amongst vaccinated individuals who experienced mortalities. Additionally, they can determine how many individuals who were infected and tested positive for COVID-19 had comorbidities.

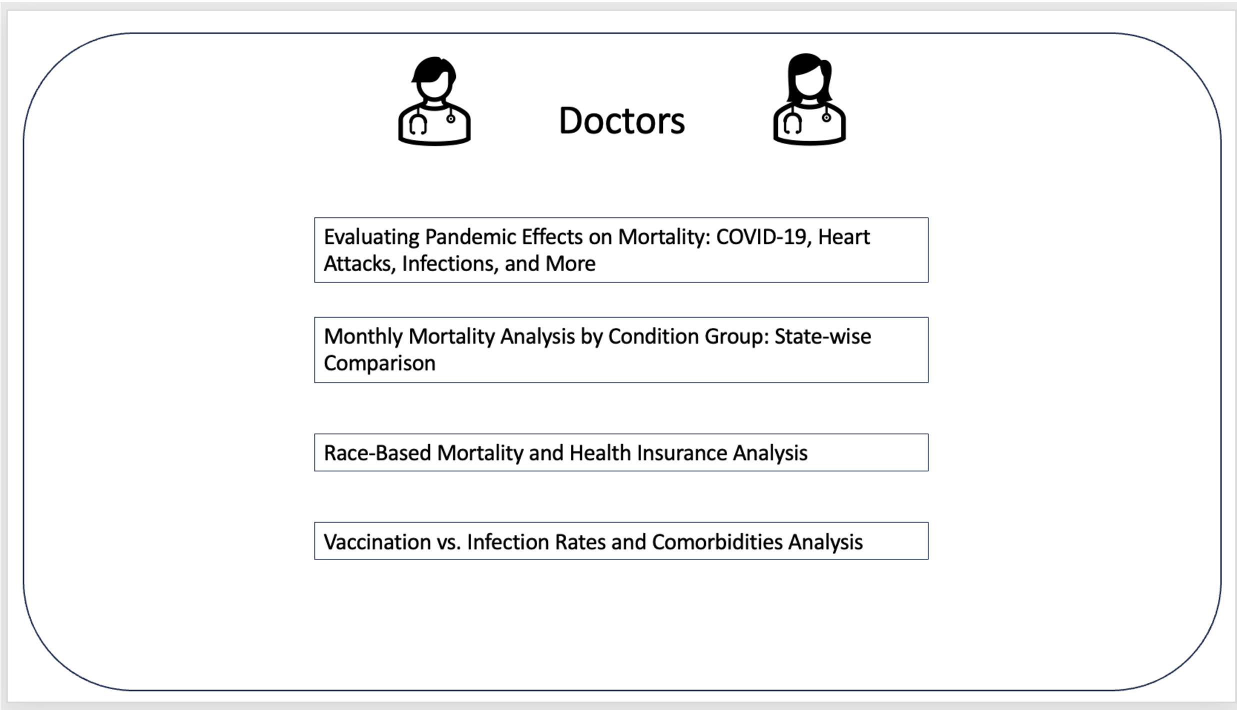
<u>The Application-Specific Network or Graph of Web Pages</u>



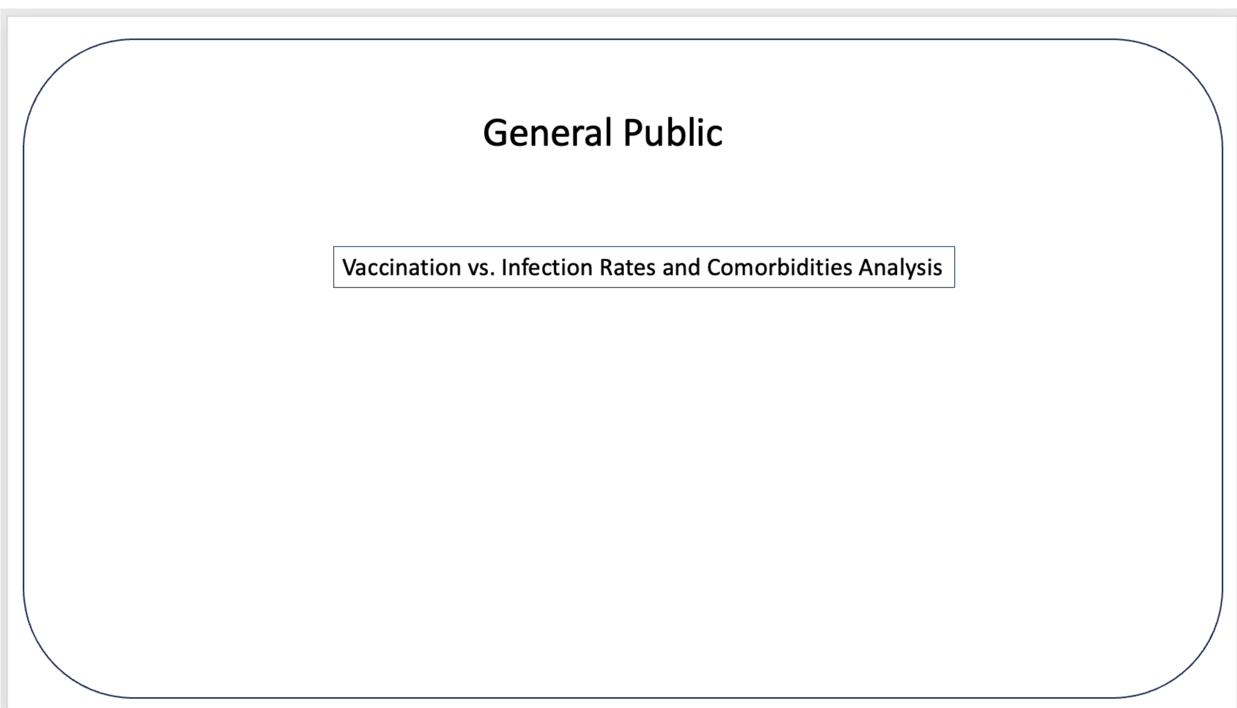
This graph represents the connectivity of the application pages to one another. The root node of the graph is the Landing page, and each child node represents a specific page which caters to a different user-base. The wireframes to better visualize these nodes are elaborated upon below:



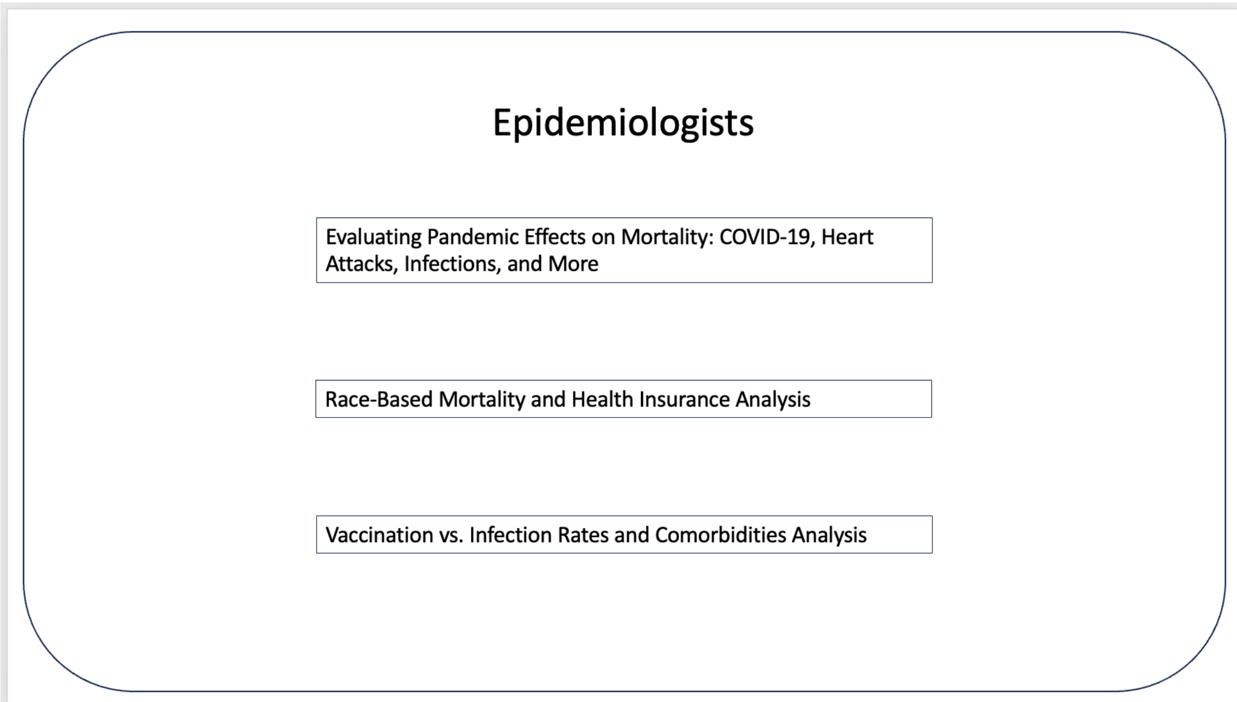
1. Landing Page: This is the landing page where users first interact with the web application. The provided icons on this page enable users to choose their area of interest, guiding them to personalized pages based on their selection.



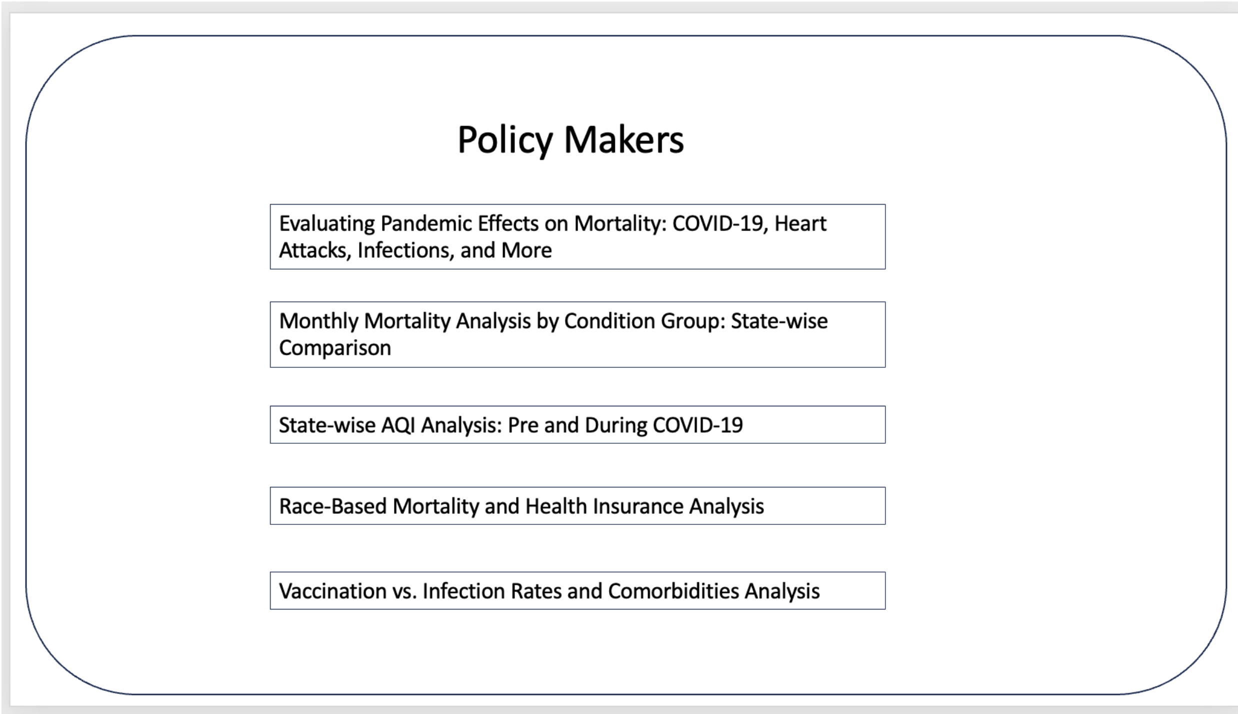
2. Medical Section (User: Doctor): The Medical section is tailored to meet the needs of doctors. It equips them with the analytical tools and datasets required to conduct in-depth retrospective analyses of the pandemic. From examining the impact of comorbidities on mortality rates to investigating the allocation of healthcare resources, this section provides a wealth of resources for professionals on the front lines. Dynamic Visualizations and complex trend queries allow for insights and informed decision-making in the healthcare field.



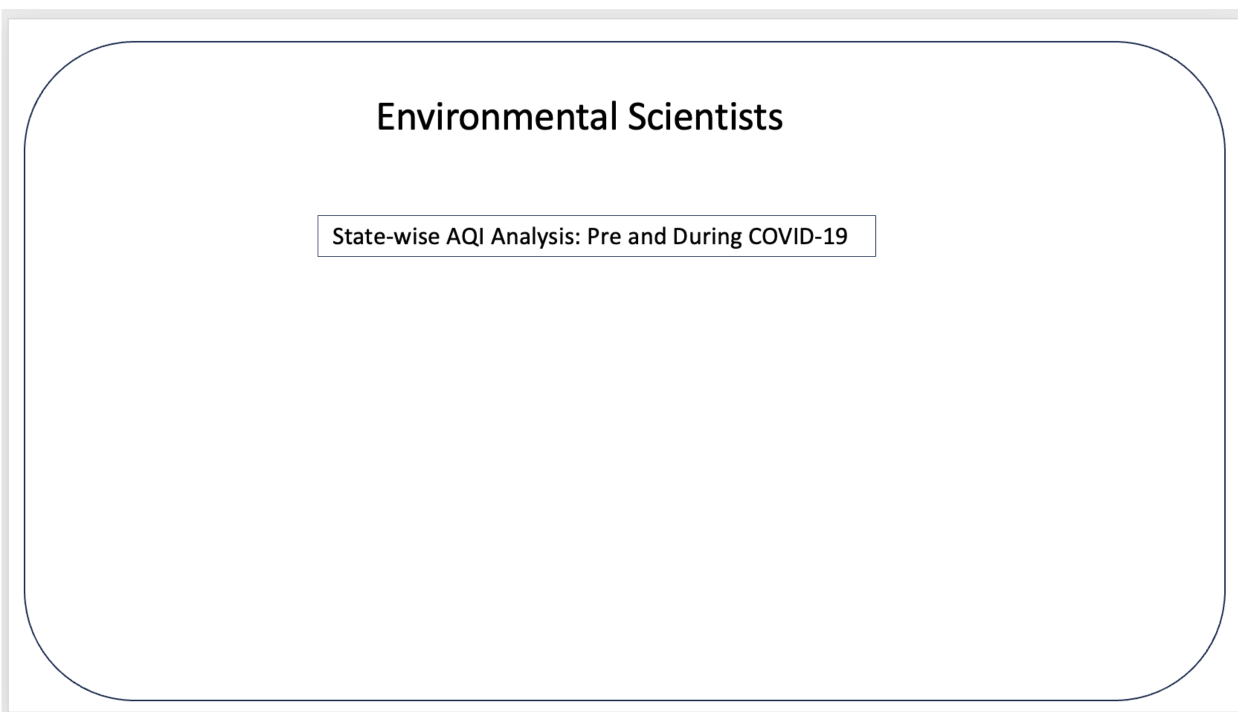
3. General Information (User: General Public): The General Public section is designed to provide everyday citizens in the United States with accessible and factual information about the COVID-19 pandemic. It offers a comprehensive view of the pandemic's progression, enabling users to explore critical data related to infection rates, vaccination coverage, healthcare capacity, and more. By offering transparent and up-to-date information, this section empowers individuals to make informed decisions, understand trends, and contribute to the collective effort to combat the virus.



4. Epidemiological Research (User: Epidemiologist): The Epidemiological Research section is meticulously designed to cater to the specific requirements of epidemiologists. It empowers them with a comprehensive suite of analytical tools and datasets, enabling in-depth exploration of COVID-19's disease dynamics. This encompasses investigating the various factors contributing to its transmission, assessing the efficacy of vaccinations and booster doses, scrutinizing the impact of comorbidities on mortality rates, and optimizing healthcare resource allocation. Within this section, professionals at the forefront of epidemiology will find a rich repository of resources. With dynamic graphs and sophisticated trend analysis capabilities, it facilitates dynamic insights and supports data-driven decision-making in the healthcare domain.



5. Policy Research (User: Policy Maker): Policymakers play a pivotal role in shaping our collective response to the pandemic. This section empowers those in positions of power to formulate and adapt policies and strategies effectively. Decision-makers can draw upon critical insights into the pandemic's various dimensions, from the socioeconomic and racial factors influencing outcomes to the environmental impact of our responses. This section serves as a vital resource for driving evidence-based policies that better the lives of constituents.



6. Environmental Science Research (User: Environmental Scientist): The Environmental Science Research section is curated to meet the unique needs of environmental scientists and researchers. This specialized area equips professionals with a comprehensive toolkit and extensive datasets to delve into the intricate connections between the COVID-19 pandemic and the environment. This section enables researchers to carry out investigations into the environmental aspects of the pandemic, specifically the air-quality index. The datasets and analytical tools provide a valuable platform for studying the pandemic's environmental dimensions and contributing to a holistic understanding of its effects on our planet.

<u>Conceptual Database Design - ER diagram</u>

Entity Relationship Model:

A. COVID-19 Dataset:

This relation contains the patient-level data U.S. states and autonomous reporting entities.

1. **<u>Patient_id</u>:** Unique and not null ID of the patient.
2. **Death_yn:** Stores Boolean value of whether the patient died or not.
3. **Case_month:** Month in which the case was reported.
4. **Underlying_conditions_yn:** Did the patient have any comorbidities or not.
5. **Race:** Race of the patient.
6. **Res_state:** State of the patient.
7. **Res_county:** County in which the patient resides.

B. Comorbidity Dataset:

This relation contains the number of patients who died due to comorbidities grouped by month and state.

1. **<u>Month</u>:** Denotes the month of passing for a specific group of individuals.
2. **<u>Year</u>:** Denotes the year of passing for a specific group of individuals.
3. **Condition:** Refers to the comorbidity the patient had.
4. **COVID-19 Death:** Stores the number of people who died in that month and had specific comorbidity.
5. **<u>State</u>:** Identifies the State for the tuple.
6. **Number of Mentions:** The number of mentions in the table represents the number of instances where the condition was cited on the death certificate.
7. **<u>ICD10</u>:** Refers to code assigned for a particular condition.

C. Health Insurance Dataset:

1. **Insured:** Number of people that have health insurance from the race.
2. **<u>Race_hispanic_origin</u>:** Identifies the race and origin (Hispanic or Non-Hispanic) of the group of people.
3. **<u>Region</u>:** Refers to the region of residence in the United States.
4. **Uninsured:** Number of people that do not have health insurance from the race in column 'Race_hispanic_origin'.

D. Daily Pollutant Concentration:

1. **Date:** This identifies the specific date for which we have the data available.
2. **Site_ID:** Unique ID assigned to a data collecting site from where we are collecting the data.
3. **Dailymax:** Maximum concentration of the pollutant for that day.

E. Vaccination:

1. **<u>Date</u>:** This identifies the specific date for which we have the data available.
2. **Series_completed:** Represents the percentage of people who have taken all required doses.
3. **Booster_doses:** Stores number of people who completed Booster doses.

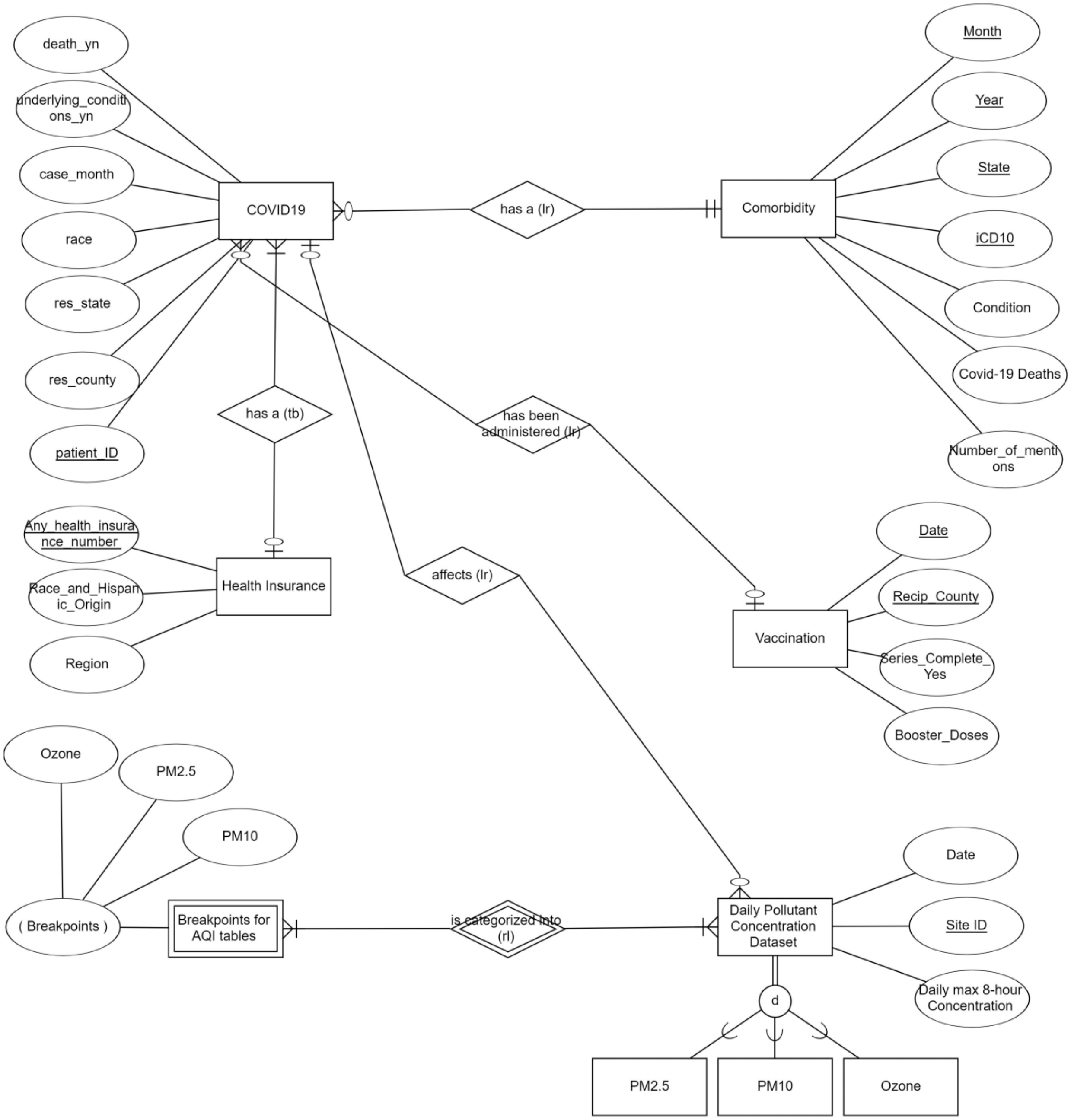


Fig: Conceptual ER diagram of the Multifaceted COVID-19 database

The ER diagram plotted above represents the various schemas and their relationships.

Mnemonics used to represent directions of relationships:

1. lr: left to right
2. rl: right to left
3. bt: bottom to top
4. tb: top to bottom

Elaboration of Relationships Displayed in the ER diagram:

1. has a (lr): This relationship between the COVID-19 schema and the Comorbidity schema represents which comorbidity (health condition) lead to the death of the COVID-19 patient. Every COVID-19 patient may or may not have died due to a comorbidity and might have died due to utmost 1 comorbidity (0:1). Also, at least 1 person would have died due to a particular comorbidity (otherwise, it won't exist in the table) and many people would have died due to that comorbidity (1:m).
2. has a (tb): This relationship between COVID-19 schema and Health Insurance schema helps us identify what health insurance plan a COVID-19 patient has. Every COVID-19 patient may or may not have insurance and can have utmost 1 insurance (0:1). Also, a particular health insurance plan may or may not be bought by any patient and can be bought by multiple patients (0:m).
3. has been administered: This relationship between COVID-19 schema and Vaccination schema tells us if a COVID-19 patient has been administered a vaccine shot or not. It also tells us how many doses of the vaccine the patient has been administered. A COVID-19 patient may or may not have been administered a vaccine shot and can be administered at max 1 type of vaccine (0:1). Also, a particular vaccine may or may not be administered to any patient or can be administered to multiple patients(0:m).
4. affects: This relationship between COVID19 schema and the Daily Pollutant Concentration Dataset schema helps us determine the trend in the concentrations of Ozone and various pollutants in the atmosphere over time. We can visualize this data to see if there was a dip in the concentration of any pollutant during the COVID19 pandemic. There may or may not be a significant rise or dip in the concentrations of ozone or any other pollutant, but there may be a change in trends of multiple pollutants (0:m). If there is a change in the trend(s) of one or more pollutants, it may or may not be due to the COVID-19 pandemic (0:1).
5. is categorized into: This relationship between Daily Pollutant Concentration Dataset schema and Breakpoints for AQI Tables schema segregates the Daily Pollutant Concentration Dataset into several categories which makes it easier to classify it into several categories from good air quality to hazardous. A particular pollutant's concentration value can fall into 1 and 1 category only (1:1). Also, a particular category may or may not have recordings of a pollutant’s concentration and can have multiple recordings (0:m).

Analysis of Deployed Data Sources

Query 1: How many people died due to other causes during the pandemic time period (cardiovascular death rate, other infectious disease deathrate)? How many COVID-19 infections were successfully treated and how many COVID-19 patients died?

Datasets: COVID-19 Dataset, Comorbidity Dataset

Columns of Interest COVID-19 Dataset: death_yn, underlying_conditions_yn, case_month
PRIMARY KEY: the data is regarding deidentified patients therefore we will be adding an extra column with a randomized unique patient ID to serve as a primary key.

Columns of Interest Comorbidity Dataset: Condition, Start_Date, End_Date, Covid-19 Deaths, Number_of_mentions
PRIMARY KEY: Month, Year, State, and iCD10 code will serve as primary key.

Note: The data from the first dataset will need to be grouped by month and the query will be performed on monthly or yearly data.

The number of mentions presented on the table above represents the number of instances where the condition was cited on the death certificate. Number of deaths and number of mentions reported in this table are tabulated from deaths received and coded as of the date of analysis and do not represent all deaths that occurred in that period. Data for this table is derived from a cut of the National Vital Statistics System (NVSS) database taken at a particular time, separate from other surveillance tables on this page which are tabulated on the date of update. As a result, the total number of COVID-19 deaths in this table may not match other surveillance tables on this page.

Query 2: What was the mortality rate of different races over time during the COVID-19 pandemic? For each region, which race had the highest mortality rate? How many people from that race had health insurance in each region vs. How many people from that race had no health insurance in each region? *Region (Northeast, Midwest, South, and West)
Datasets: COVID-19, Health Insurance

Columns of Interest COVID-19 Dataset: race, res_state, death_yn

PRIMARY KEY: the data is regarding deidentified patients therefore we will be adding an extra column with a randomized unique patient ID to serve as a primary key.

Columns of Interest Health Insurance Dataset: Any_health_insurance_number, Race_and_Hispanic_Origin, Region

Query 3: Find the mean AQI for each pollutant over a period of time for each state. Following this compare the AQI data points (for a year prior to COVID-19 with a year during the pandemic, ex: 2018 vs. 2020) with the Mann-Whitney U-test to determine whether the points are significantly different or not?

Dataset: Daily max 8-hour Concentration datasets for different pollutants,
Columns of Interest: Date, Site ID, Daily max 8-hour Concentration.
Dataset: Breakpoints for AQI table
Columns of Interest: Breakpoints

Query 4: Over time which condition group (ex. circulatory diseases, respiratory diseases), in each state, led to how many mean mortalities per month? Compare these against one another using the One-Way ANOVA test and determine what is the difference in mortality with reference to the other condition afflicting the patients?

Datasets: COVID-19 Dataset, Comorbidity Dataset
COVID-19 Dataset:
Columns of Interest: Number_of_Mentions, res_state, res_country, death_yn, underlying_condition_yn
Comorbidity Dataset:
Columns of Interest: Month, Year, State, Condition, COVID-19 Deaths, Number of Mentions

Note: Number of deaths and number of mentions reported in this table are tabulated from deaths received and coded as of the date of analysis and do not represent all deaths that occurred in that period. Data for this table is derived from a cut of the National Vital Statistics System (NVSS) database taken at a particular time, separate from other surveillance tables on this page which are tabulated on the date of update. As a result, the total number of COVID-19 deaths in this table may not match other surveillance tables on this page.

Query 5: What was the vaccination rate vs. infection rate over time? In the mortalities amongst people who were vaccinated did they have comorbidities? How many people who were infected and tested positive had comorbidities?

Datasets: COVID-19 dataset, Vaccination dataset

Columns of Interest in COVID-19: underlying_exposure_yn, conditons_yn, case_month, current_status, death_yn
Columns of Interest in Vaccination Dataset: Date, Series_Complete_Yes, Administered_Dose1_Recip, Booster_Doses