# Research Paper Midway Checkpoint

## Summary

This project investigates how socioeconomic and demographic factors influenced telehealth adoption patterns in the United States during the COVID-19 pandemic (2020-2022), focusing on racial and ethnic disparities, geographic differences, and educational attainment.

Using publicly available datasets from RANDS and Medicare Telehealth Trends, the study employs temporal and comparative analyses to quantify disparities, identify barriers to access, and provide evidence-based policy recommendations for equitable healthcare delivery.

The research aims to highlight how existing healthcare inequities were either mitigated or exacerbated by the rapid expansion of telehealth services.

## Research Question

We will explore the following Research Question through this project:

**How did socioeconomic and demographic factors influence telehealth adoption patterns during the COVID-19 pandemic in the United States (2020-2022)**

The impact on the telehealth adoption patterns will be explored via the following factors:

1. ***Racial*** and ***ethnic disparities*** in telehealth utilization rates
2. ***Geographic differences*** in adoption patterns (privileged and under-privileged)
3. The relationship between ***educational attainment*** and telehealth usage

Our Research Question did not change as we had a discussion with TA about it being properly scoped and also how our datasets can further answer the question. There was no extra information provided in the feedback which we assumed to mean that the Research Question is well formulated and appropriately scoped given the data. Thus, the Research Question did not change for us.

## Data cleaning

We used a jupyter notebook to clean, analyse and visualise (EDA) the data. The reasoning was that Jupyter Notebooks would allow us to interleave pieces of text between the code for better communication within the team as well as to the TA.

The link to the code (Also submitted on Canvas):

We first loaded the data into the Jupyter Notebook. We are cleaning the data before uploading it to SQL and combining, this is done to so, we can define it for a proper schema.

For the ***“Access\_and\_Use\_of\_Telemedicine\_During\_COVID-19.csv***”:

1. Remove all the irrelevant columns such as “Suppression, Significant 1 and Significant 2” reasoning for which was given in Research proposal.
2. Look at all the Summary statistics for the numerical rows:
   1. The summary statistics for all the numerical rows make sense as they are non-zero and meet expectations. This is done by comparing it to the values seen in the csv and don’t indicate any major outliers.
3. Look at all the NULL values in the rows:
   1. As all the rows have the group and sub-group listed, all of them are useful. Even if they are `Null`, they are providing certain information about the groups and subgroups that is useful
4. Look at all the Sub-groups of the groups interested
   1. Races primarily focus on Hispanic and non-Hispanic ethnic disparities. The Hispanic race is further broken down into White and Black Non-Hispanic. This could provide further challenges when matching the dataset when merging with the other dataset. This will be further discussed when analysing the other dataset.
   2. We will convert ‘Metropolitan’ and ‘Non-Metropolitan’ to “Urban” and “Rural” respectively, due to the reasoning mentioned in the Research Proposal: In our study, we will create consistent geographic units for rural/urban classification: Our assumption will classify “metropolitan” cities and “urban” areas as one classification which will represent more connected and privileged residential areas. Similarly, we will classify “non-metropolitan” and “rural” areas as one classification to represent less connected and generally under / less privileged residential areas.
   3. We will also convert 'High school graduate or less' to ‘highschool or less’, 'Some college' to ‘college’ and "Bachelor's degree or above" to ‘bachelor or above’ for convenience.
   4. We will also ignore the “Urbanisation” (different from point b) and other subgroups besides the one mentioned above as it does not provide any useful information or any further subgrouping required for our research question.
5. Finally, add a column called “size” that shows not only the size instead of the percentage of the population which will make further calculation easier. This will be calculated per SubGroup of the relevant Groups per Indicator.
   1. We first calculated the Total Sample Size per ‘Indicator’ per ‘Group’ and then mapped it to each row accordingly then the Percentage and Total Response were used to calculate the data.

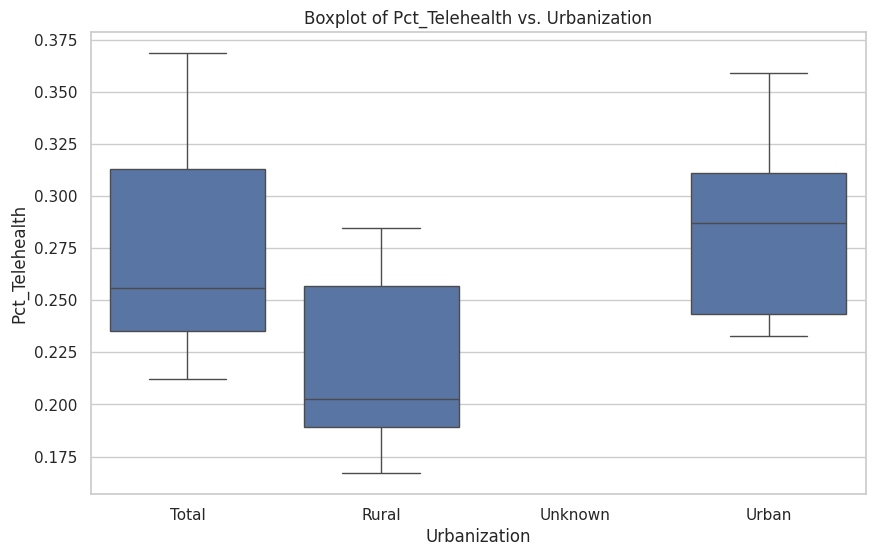
For the “***TIMEDTREND\_PUBLIC\_241126.csv***”:

1. We first filter only the COVID years which is 2020-2022 (inclusive)
2. Look at all the Summary statistics for the numerical rows:
   1. The summary statistics for all the numerical rows make sense as they are non-zero and meet expectations. This is done by comparing it to the values seen in the csv and don’t indicate any major outliers.
3. Look at all the NULL values in the rows:
   1. Almost all columns have complete data and are not NULL so no rows are removed.
4. We then aggregate the data for our relevant purposes which include ‘Bene\_Race\_Desc’, ‘Bene\_RUCA\_Desc’, ‘Bene\_Mdcd\_Mdcr\_Enrl\_Stus’ for each numerical column. While aggregating, every absolute value is taken as a sum except percentage which is taken as a mean as the summation would not make much sense for a relative metric.
5. We also renamed the column to ‘Race’, ‘Urbanization’, and ‘Enrollement\_Status’ for ease of understanding.

We will use Urbanisation and Race to combine the datasets, so, we will convert the attribute values to match with the other dataset to keep it consistent. This will also make the SQL Script for data design easier:

* ‘All’ to ‘Total’ for both ‘Race’ and ‘Urban’
* 'Non-Hispanic White' will be renamed to ‘White Non-Hispanic’
* 'Black/African American’ will be renamed to ‘Black Non-Hispanic’
* ‘Other/Unknown’ will be renamed to ‘Other Non-Hispanic’
* ‘Hispanic’, ‘Urban’, ‘Rural’ will stay the same as they are the same in the other dataset.
* The rest of the values in Race and Urbanisation do not match, but will remain the same as only the cleaned version of the first dataset (“***A*ccess\_and\_Use\_of\_Telemedicine\_During\_COVID-19.csv**”) will be referencing the cleaned version of the second dataset (“***TMEDTREND\_PUBLIC\_241126.csv***”). This will be further clarified in the *7. SQL Schema step*.

## EDA

1. 
2. The boxplot visualizes the distribution of the Pct\_Telehealth variable (percentage of telehealth usage) across different levels of the Urbanization variable. It shows how the distribution of telehealth usage differs between urban and rural areas, as represented in the df2\_final dataset. Specifically, it displays the median, quartiles, and potential outliers for telehealth percentage in each urbanization category. This allows for a quick comparison of variability in telehealth usage patterns between urban and rural populations.
3. The boxplot of 'Pct\_Telehealth' against 'Urbanization' in df2\_final provides a preliminary view. The following can be done needs to be done after the exploratory analysis:

- Conduct statistical tests (possibly t-test) to determine if the difference in telehealth usage across urbanization levels is statistically significant.

- Explore potential interactions between urbanization and other variables (e.g., Race, Enrollment Status) on telehealth usage.

**-** Investigate "Unknown" Category: Check what contributes to the "Unknown" urbanization category and whether it should be excluded or recategorized.

## EDA effect

The EDA results suggest that urbanization level significantly influences telehealth usage, which affects how the data should be handled moving forward. The regression model used for the temporal analysis will have to include the Urbanization subgroups, although this would need to be verified using a t-test. As there is overlap between the different levels, it would be sufficient to use One-Hot Encoding for the subgroups, rather than treating them as separate covariates. If EDA had shown no clear difference in telehealth usage across urbanization levels, there wouldn’t be a need to adjust the handling of this variable—possibly treating it as a non-informative feature. However, the visible differences indicate that urbanization is an important factor.

## SQL Script

As there are too many INSERT statements per value, we generated the SQL INSERT statements using Python code which is present in the submitted code file. The SQL code itself is too long to be listed in the current document but is present in the attached .sql file.

## SQL Schema

For the cleaned version of the ***“Access\_and\_Use\_of\_Telemedicine\_During\_COVID-19.csv***” dataset:

**CREATE** **TABLE** TelemedicineProvider (

Round INT,

**Indicator** VARCHAR(255),

**Group** VARCHAR(255),

SubGroup VARCHAR(255),

SampleSize FLOAT,

Response VARCHAR(255),

**Percent** FLOAT,

StandardError FLOAT,

Size FLOAT,

**PRIMARY** KEY (Round, **Indicator**, **Group**, SubGroup, Response),

);

For the cleaned version of the “***TIMEDTREND\_PUBLIC\_241126.csv***” dataset file:

**CREATE** **TABLE** BeneficiaryData (

Race VARCHAR(255),

Urbanization VARCHAR(255),

Enrollment\_Status VARCHAR(255),

Total\_Bene\_TH\_Elig FLOAT,

Total\_PartB\_Enrl FLOAT,

Total\_Bene\_Telehealth FLOAT,

Pct\_Telehealth FLOAT,

**PRIMARY** KEY (Race, Urbanization, Enrollment\_Status)

);

In the context of the provided datasets, using Group and SubGroup as foreign keys in the TelemedicineProvider table may not be appropriate because not all values in these columns will necessarily match the corresponding values in the BeneficiaryData. Foreign keys enforce referential integrity, meaning that every value in the Group and SubGroup columns must exist in the referenced tables. However, we do not want to filter out or exclude certain Group and SubGroup values that do not align with the analysis criteria. This filtering process would result in some rows being excluded from the joined datasets, which contradicts the strict enforcement of foreign key constraints. Therefore, it is more practical to handle these relationships during the analysis phase rather than enforcing them as foreign keys in the database schema. This approach allows for greater flexibility in data manipulation and ensures that the analysis can proceed without being constrained by referential integrity rules that do not align with the analytical goals.

There are no other constraints besides the PRIMARY KEY listed above as most of the NOT NULL requirements are met by the PRIMARY KEY itself. The other attributes do not have other requirements listed.

## AI declaration

We have not used any AI or GenAI tools in this research project so far, All the steps have been completed with the materials learned in class or researched online (Links provided above).

* Aarav, Pushya, Suryansh