Let's Get Clinical: Racial and Ethnic Representation in Clinical Trials for FDA-Approved Drugs

Team Members

- Kristy Kwon (kkwon35)
- Caitlin Pratt (caitlinp)
- Alison Spencer (alisonspencer)
- David Steffen (davidrsteffen)

Abstract

In our project, we will analyze representation in clinical trials for different treatments and identify trends over time. First, we collected data on approved treatments from the FDA and on clinical trials from the ClinicalTrials.gov website. Then, we cleaned the datasets, merged them in order to identify the trials conducted to test each approved treatment, and analyzed them in order to determine the racial and ethnic representation in each trial. Finally, we created graphical visualizations and tables of racial and ethnic representation in the trials for particular treatments and in the trials sponsored by particular manufacturers and presented them in an interactive dashboard.

Pharmaceutical companies use clinical trials to test the effectiveness of their products. Clinical trials are required by the U.S. Food and Drug Administration (FDA) before new treatments are approved for use, and their results are routinely used to determine which treatments patients should take to treat their medical conditions.

Introduction

Over the last two decades, the FDA has approved a growing number of health-related drugs and treatments, a rate that is increasing at approximately 12.5 percent annually on average. These include around 11.96 to 17.54 percent of anti-cancer drugs, 9.09 percent of remedies for cardiovascular conditions, and 7.17 to 15.96 percent of vaccines and antibodies. These treatments are vital for many Americans who suffer from various ailments. For example, 1 in 5 Americans died from cardiovascular diseases or attacks in 2021, 50 to 55 percent of adults seek vaccines for seasonal colds, and the estimated cancer incidence per year is 442.4 per 100,000 men and women in the U.S.

As such, the clinical trials implemented by pharmaceutical companies are crucial to ensure the effectiveness of medicinal remedies and determine the appropriate drugs for patients. However, many clinical trials do not enroll subjects who accurately reflect the general population. Often, these studies are too limited in scope and time period, and neglect to control for baseline characteristics to make adequate comparisons across individuals. Therefore, pharmaceutical

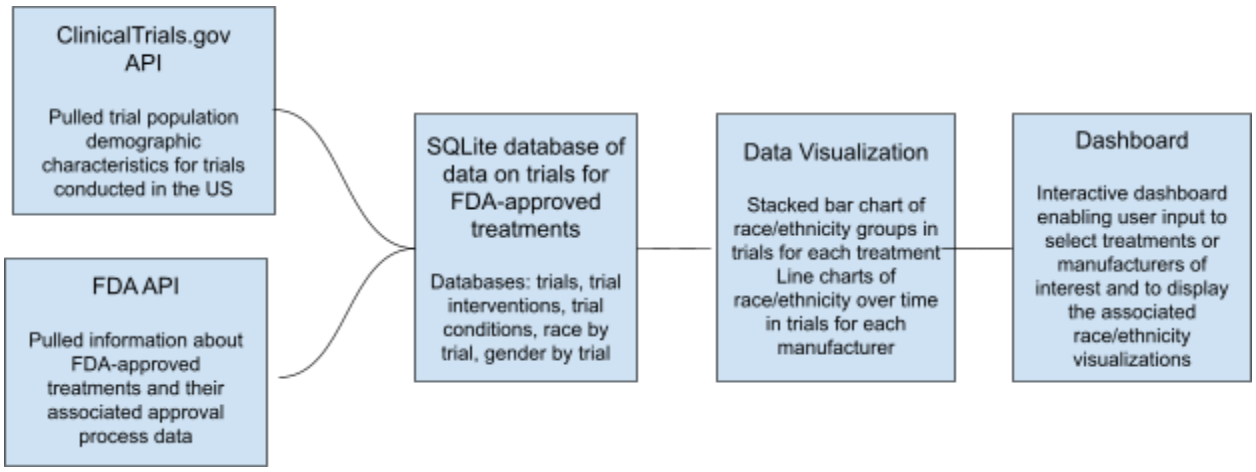
companies are unable to produce treatments that can consistently work and serve clients of all backgrounds.

In particular, there has been frequent underrepresentation of racial and ethnic minority groups in clinical trials. A study in 2020 indicates the dearth of members from marginalized communities: among 30,000 clinical trial participants, 6 percent were Asian, 8 percent were Black, and 11 percent were Hispanic. Historically, these groups have had poorer health outcomes relative to the overall population, exacerbated by the lack of access to systems and resources to address basic health needs. For instance, the infant mortality rate is two times higher for Black Americans compared to the national average of 5.8 births in 100,000, and Hispanic Americans are 66 percent more likely to have type 2 diabetes. The public has pushed for pharmaceutical companies to address these disparities through changes in how they select participants to test their products. Affirming this importance, over 500 pharmaceutical organizations have recently committed to increasing diversity in clinical trials.

Analysis of racial representation in the clinical trials data can shed light on the extent to which companies have been successful in accounting for different identities and characteristics in their dataset over time.

Overall Software Structure

Project Module Diagram



Code Responsibilities

Data Collection

- **Caitlin** created the function to pull data from the ClinicalTrials.gov API
- **Alison** created the function to pull data from the FDA API

Database

- **Caitlin and Alison** worked together to design and create the SQL database
- **Caitlin** formatted the ClinicalTrials.gov data for inclusion in the database
- **Alison** formatted the FDA data for inclusion in the database

Cleaning and Analysis

- **Caitlin and Alison** worked together to filter the data to trials of interest that were for Phase III approved trials, conducted in the U.S., for drugs that are now FDA approved.
- **Caitlin** worked on cleaning the race and ethnicity data for each trial in the ClinicalTrials.gov API data, and joining the trial sponsor data to the FDA manufacturer data
- **Alison** worked on cleaning the incoming FDA data
- **Kristy** analyzed the data to display the racial and ethnic representation in each clinical trial, and generated summary statistics for each trial and each manufacturer

Data Visualization

- **Kristy** wrote the functions to create the graphs of interest
- **Kristy** wrote the functions to generate the summary statistics of interest

Dashboard

- **David** designed the dashboard, integrated the SQL database into the dashboard, and integrated the visualizations and tables into the dashboard
- **David and Caitlin** worked together to design queries to pull data from the SQL database for each treatment and each manufacturer
- **David and Kristy** worked together to integrate the graphs and tables into the dashboard

Admin

- **Caitlin** created the Poetry virtual environment and the GitHub repository, and was our resource for GitHub issues and questions
- **Caitlin** set up Main so the project can be run from the command line
- **Caitlin** wrote the READ-ME.md
- **David** wrote the abstract and code responsibilities sections of the proj-paper.pdf
- **Kristy** put together the intro/rationale, findings, and conclusion in proj-paper.pdf
- **Alison** contributed to the goals section of pro-paper.pdf.

Application User Guide

The application first collects data from the ClinicalTrials.gov API and FDA API. This data is cleaned into csv files and loaded into a sqlite3 database to access for data visualization and building the dashboard.

To get this data:

Neither the FDA api nor the NIH Clinical Trials API require an API key. Pulling should be possible out of the box. JSON data is saved to the /data parent directory mentioned in the setting up section.

1. Fetch the FDA api data. This data is the core dataset that we will use to validate drug names in the trials data. These pulls should take about 2 minutes each to run.

    Run: python3 clinicaltrials/api/fetch_fda_data.py

2. Next, fetch the NIH clinical trials api data:

    Run: python3 clinicaltrials/api/fetch_trials_data.py

    After running this, you should see that the /data folder in the parent directory now includes two files: fda.json and trials.json

3. Extracting API data: Once API data is pulled, it should not be necessary to pull again. Whenever API data is pulled, it should be extracted and cleaned. Do this by running the following commands:

    python3 clinicaltrials/data/extract_fda_data.py
    python3 clinicaltrials/data/extract_trials_data.py

    Note the following files in the same directory as the extraction scripts:

    clinicaltrials/dedupe_dataframe_learned_settings
    clinicaltrials/dedupe_dataframe_training.json

    These are trained classifier files used to perform fuzzy deduplication of FDA drug records. Do not delete them. If you do, you will have to retrain the classifier.

4. Once you have run the cleaning screepts, you should see that the csvs parent data directory is now populated with nine CSVs (fda_full.csv, trial_conditions.csv, trial_interventions_raw.csv, trial_interventions.csv, trial_locations.csv, trial_race.csv, trial_sex.csv, and trial_status.csv trials.csv

Note that while data about sex representation in trial is extracted, the current version of the app presently does not display data on the sex of trial participants. It is the hope of the clinical-trials team to continue maintaining this tool, and to incorporate this and further demographic analysis after the project is submitted.

5. Populating the local database: Once data has been extracted and saved to csvs, run the script to populate the local database:

    python3 clinicaltrials/data/makedb.py

6. Running the app: At this point, everything you need to run the app should be there. Run the app with the following command, and input the address into your browser to view the tool:

    python3 clinicaltrials/app.py

The dashboard has six components:

1. **Dropdown Menu 1:** Presents the generic name, brand name, sponsor, and approved date of the treatment of interest

2. **Stacked Bar Chart:** Visualizes racial breakdown of participants in each clinical trial by condition and treatment of interest

3. **Summary Statistics 1:** Displays the mean, maximum, median, range, and interquartile range of participants for certain race categories by condition and treatment of interest, as well as the overall total and average number of observations

4. **Dropdown Menu 2:** Lists the manufacturer name and range of approved dates for each clinical trial by manufacturer

5. **Line Graph:** Examines the average number of participants in a given race by each manufacturer from 2009 to 2023

6. **Summary Statistics 2:** Shows the mean, maximum, median, range, and interquartile range of participants for certain race categories by manufacturer of interest, as well as the overall total and average number of observations

Findings and Goals

In the aggregate, both the line graph based on the manufacturer as well as the stacked bar chart by treatment and condition reveal that non-white participants were represented by at least a factor of two times to at most a factor of seven to eight times less than their white

counterparts. Compared to the filtered data by treatment and condition, the data subsetted by manufacturers indicate overrepresentation of white participants relative to their non-white counterparts.

Several treatments and conditions either consisted only of white participants or included a single racial minority group, which only consisted of 10 to 20 percent or less of all participants in the clinical trials. For other remedies and ailments, while there were at least two to three non-white participants represented, each of them accounted for at most roughly 1.5 times to two times less than the number of white participants.

From 2009 to 2023, on average, white participants make up 70 to 90 percent of the share of any participants in clinical trials by manufacturer. On the other hand, non-white participants comprise 0 to 20 percent of the share of total participants in clinical trials by manufacturer. From this time period, the rate of representation for black participants by manufacturers fluctuates from 6 percent to as high as 20 percent. This rate is around a 5 to 15 percent range for asian participants by manufacturers, approximately 0 to 5 percent for hispanic participants, and 0 to 4 percent for participants of other ethnicities. We conclude that white participants were overrepresented while minority participants were underrepresented.

This project set out to examine racial and ethnic diversity in clinical trials, by treatment and by drug manufacturer, and to examine how this diversity has changed over time. We were able to accomplish this through combining data from ClinicalTrials.gov and the FDA. We organized this data in a way that people can see visualizations of this issue. Racial diversity in clinical trials has received increased attention in recent years. One issue is that collecting data on the lack of diversity in clinical trials can be difficult, due to trials sometimes not reporting race, or to trials reporting race differently. This project aimed to organize racial and ethnic categorization of clinical trials across various trials in a way that it was possible to generate summaries and observe general trends over time. The linking of the clinical trials data to FDA data provides a way to see what trials were conducted specifically for approved drugs.

Linking to the FDA data also provides a way for application users to see not only how clinical trials for the drug were conducted, but also other aspects of the FDA approval process. We provide submission status date and FDA application number so that more FDA information on the drug can be seen. By using the ClinicalTrials.gov and FDA APIs, we have also provided a way for this information to be updated in the future. Data could be pulled from these sources at a later date, and the application could be updated correspondingly. In this way, we could continue to see trends in drug trials and if diversity improves.