Chronic Cough

- By Team Data Wranglers

Project repository link: https://github.com/MeghnaMSCS/BMI-6016-Chronic-Cough

Alex Kimball | alejandro.kimball@utah.edu | u1195729
Jenine Rogel | J.Rogel@utah.edu | u0468294
Joseph Worden | joseph.worden@hsc.utah.edu | u0919070
Meghna Manjunatha | u1368460@utah.edu | u1368460
Siwei Zhou | u1421984@utah.edu | u1421984

1. Motivation

Our motivation for the project stems from personal experiences of our team mates. This is what they have to say:

“ After catching COVID at the end of the year 2020, I lost my sense of smell. A week later, I was able to smell, but my cough never went away. It was very unpleasant to have a cough when COVID was at its peak. Once, I was sent home from work awaiting COVID results. I walked into work meetings with a handful of cough drops and sat in the back. After going to the doctor at least a dozen times and trying unnecessary expensive treatments, I was finally referred to the chronic cough clinic. I felt like the diagnostic process was slow and rough. In addition, most doctors don’t understand chronic cough and need an easier way to diagnose chronic cough “

We chose this project to learn more about chronic cough, and we want to characterize chronic cough using data from TriNetX.

2. Project Objectives

The goal of the project is to characterize the phenotype of chronic cough.

2.1. We want to be able to answer the following questions.

- Does the frequency of coughing increase or decrease at different points in the year?
- Understand who develops this condition of chronic cough?
- Is there a health disparity related to chronic cough? I.E. Does it affect a greater proportion of individuals based on race, region, or socioeconomic status?

We believe the answers to these questions would help doctors better understand chronic cough and to diagnose their patients.

2.2. Benefits of how the data could be useful.
Doctors could offer better treatments if they know if the symptoms get worse in certain times of the year, and in the patients with certain underlying health conditions.

3. Data

The data came from TriNetX, https://trinetx.com/ which consists of de-identified electronic health records. There are a total of 22 files for our project which includes both .csv files and descriptive documents.

3.1 Data Processing

We are expecting to do substantial data cleaning. We have 18 .csv files that we need to review.

3.2 Must-Have Features

The essential features include: [table]Diagnosis: patient_id, encounter_id, code_system, code, admitting_diagnosis, reason_for_visit, date; [table]Patient Demographic: sex, race, ethnicity, patient_regional_location, and year_of_birth. These features will allow us to complete the project objectives which include identifying the frequency of chronic cough throughout the year, identifying who is affected by chronic cough, and identifying the phenotypes associated with a diagnosis of chronic cough.

3.3 Optional Features

If time permits, we want to analyze the following features: [table]Vital Sign: patient_id, encounter_id, code_system, code, date, value, text_value, units_of_measure, derived_by_TriNetX, source_id; [table]Patient Demographic: marital_status, reason_yob_missing, death_date_source_id, month_year_death; [table]Medication: patient_id, encounter_id, unique_id, code_system, code, start_date, route, brand, strength, derived_by_TriNetX, source_id. These features will allow us to pursue additional analysis on vital sign trends, additional demographic trends, and treatments often utilized.

4. Data Processing

We are expecting to perform substantial data clean up.
Quantities expected to derive from data in regards to chronic cough characterization:

- Structured coded data (i.e. ICD code for "Cough" being present)
- Prescriptions used specifically for cough
- # of cough encounters in a specified time period

Natural Language Processing may be implemented to catch patients with coughing that are coded for other diagnoses instead.

Data Elements Required:

- Does the frequency of coughing increase or decrease at different points in the year? [table]Diagnosis: patient_id, encounter_id, code_system, code, admitting_diagnosis, reason_for_visit, date
- Understand who develops this condition of chronic cough? Is there a health disparity related to chronic cough? I.E. Does it affect a greater proportion of individuals based on race, region, or socioeconomic status? [table]Patient Demographic: patient_id, sex, race, ethnicity, marital_status, patient_regional_locatio, year_of_birth, reason_yob_missing, death_date_source_id, month_year_death.
- Characterized this phenotype of chronic cough? [table]Diagnosis: patient_id, encounter_id, code_system, code, admitting_diagnosis, reason_for_visit, date

Data Processing Method

- Python - programming tool

5. Design

We are looking to represent the data that best answers our objectives for the project by visual representations. We would like to use Table1 format for representing our data elements (variables) since it gives a good summary of the entire data in a comprehensive way. Below are our choices for representing the results, for each of our objective questions:

- We will utilize a line graph to display the frequency of chronic cough at different time points in the year.
- We will make use of Bar Graph to display who develops Chronic cough, and to compare among the greater proportion of individuals based on race, region, or socioeconomic status.
- We will make use of heat maps, scatter plots, and spider-web graphs to characterize this phenotype of chronic cough.

6. Timeline

We will set the timeline based on feedback from our initial meeting with our instructors.

Week 1: 2/20 - 2/24

- Complete & Submit Proposal by 2/22
    - Define research questions
    - Identify Data Process Method
    - Identify Data Design
    - Clarify must have and optional features - Joe will complete
    - Set project schedule for 10 weeks which includes various steps for data processing methods, meetings with professors, team meetings, and steps for completing various project items such as presentations or a project wrap up report.
- Review existing data
    - Assign data cleaning tasks to team members
    - Finalize data cleaning plan and timeline

Week 2: 2/27 - 3/3

- Organize the dataset
- Identify key variables and potential outliers
- Determine appropriate model for analysis

Week 3: 3/6 - 3/10

- Decide specific steps for data visualization
- Assign data visualization tasks to team members

Week 4: 3/13 - 3/17

Week 5: 3/20 - 3/24

Week 6: 3/27 - 3/31

- Collect information to fill in the report

Week 7: 4/3 - 4/7

Week 8: 4/10 - 4/14

- Complete final version of the project

Week 9: 4/17 - 4/21

- Prepare presentation

Week 10: 4/24 - 4/27

- Rehearse presentation