Long-Haul COVID

Shad Morton, Logan Greydanus, John Arnn, Kevin Yang

Project Objectives

- Describe, Characterize, and Predict

Original:

- Which patients developed long-term symptoms?
- What are the symptoms and what is the severity?
- How to use these data to predict when a patient will develop long-term symptoms?

Comment on this paper

Challenges in defining Long COVID: Striking differences across literature, Electronic Health Records, and patient-reported information

Halie M. Rando, Tellen D. Bennett, James Brian Byrd, Carolyn Bramante, Tiffany J. Callahan, Christopher G. Chute, Hannah E. Davis, Rachel Deer, Joel Gagnier, Farrukh M Koraishy, Feifan Liu, Julie A. McMurry, Richard A. Moffitt, Emily R. Pfaff, Justin T. Reese, Rose Relevo, Peter N. Robinson, Joel H. Saltz, Anthony Solomonides, Anupam Sule, Umit Topaloglu, Melissa A. Haendel

doi: https://doi.org/10.1101/2021.03.20.21253896

This article is a preprint and has not been peer-reviewed [what does this mean?]. It reports new medical research that has yet to be evaluated and so should *not* be used to guide clinical practice.

Abstract Full Text Info/History Metrics Preview PDF

Abstract

Since late 2019, the novel coronavirus SARS-CoV-2 has introduced a wide array of health challenges globally. In addition to a complex acute presentation that can affect multiple organ systems, increasing evidence points to long-term sequelae being common and impactful. The worldwide scientific community is forging ahead to characterize a wide range of outcomes associated with SARS-CoV-2 infection; however the underlying assumptions in these studies have varied so widely that the resulting data are difficult to compareFormal definitions are

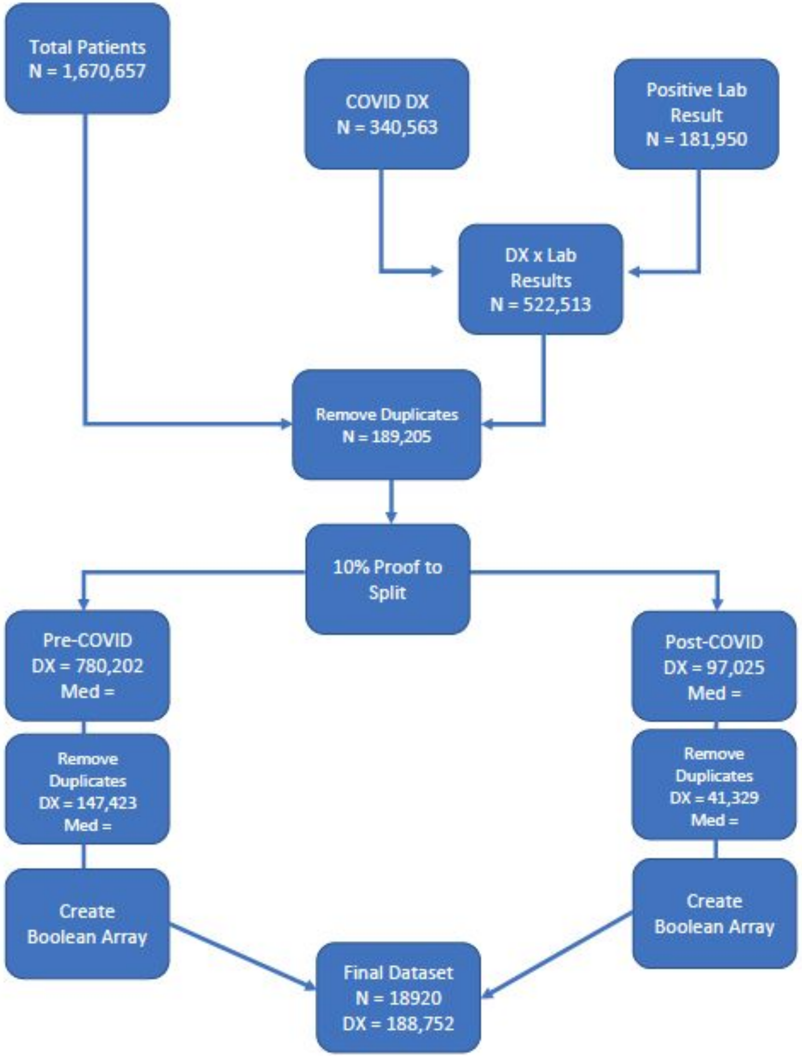
Modified:

- Prepare most common diagnosis and medication codes 14 days after Covid diagnosis for purpose of a clustering exercise.
- Followed by an analysis of clusters for their demographics and all pre-Covid diagnosis and medication codes.

Workflow

- Preform Data Assessment, Cleaning, and Profiling of original data
  - patient.csv, diagnosis.csv, medications.csv

- Integrate Data based upon previous results and goals
  - Top 200 most common diagnosis and medication codes
- Transform Data for cluster analysis
  - Boolean Array



Identifying symptoms of long-haul

- A study of symptoms of long-haul COVID from health care workers.
- This means people who had generally milder cases of COVID than other studies have investigated (non-hospitalized).
- Included 323 sero-positive participants.
- The most common symptoms were anosmia, fatigue, ageusia, and dyspnea. Only fatigue and dyspnea are present in our diagnosis dataset.
- To mirror this study, our dataset could include a control group (1.6 million - 189,000), scrape symptoms from notes based on encounter ID data, and filter procedures for intubation to assess severity of infection.

Research Letter

ONLINE FIRST FREE

April 7, 2021

Symptoms and Functional Impairment Assessed 8 Months After Mild COVID-19 Among Health Care Workers

Sebastian Havervall, MD[1]; Axel Rosell, MD[1]; Mia Phillipson, PhD[2]; et al

» Author Affiliations | Article Information

JAMA. Published online April 7, 2021. doi:10.1001/jama.2021.5612

Top 20 diagnosis codes, pre and post COVID diagnosis/positive lab result

Cough
Chest pain, unspecified
Obesity, unspecified
Other long term (current) drug therapy
Shortness of breath
Anemia, unspecified
COUGH AND COLD PREPARATIONS
Urinary tract infection, site not specified
Hyperlipidemia, unspecified
Encounter for general adult medical examination without abnormal findings
Gastro-esophageal reflux disease without esophagitis
Other chronic pain
Unspecified abdominal pain
Type 2 diabetes mellitus without complications
Fever, unspecified
Acute upper respiratory infection, unspecified
Cervicalgia
Headache
Personal history of nicotine dependence
Low back pain

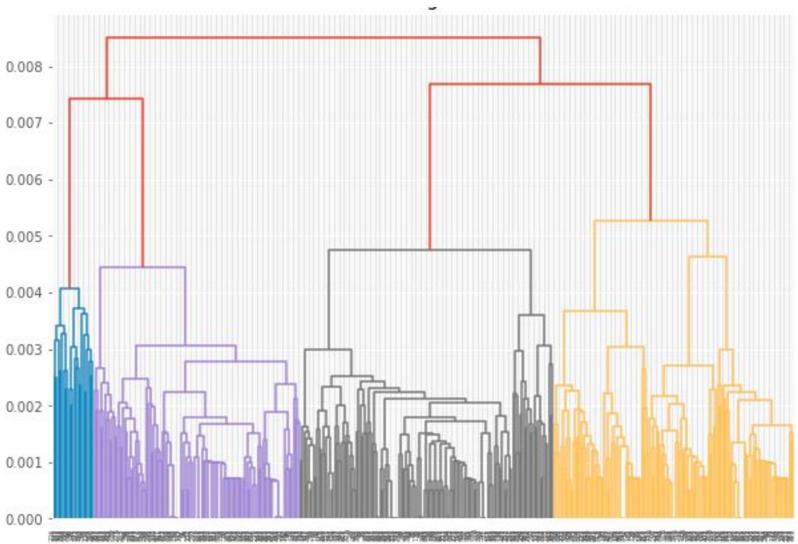
Cough
Shortness of breath
Other long term (current) drug therapy
Hyperlipidemia, unspecified
COUGH AND COLD PREPARATIONS
Acute respiratory failure with hypoxia
Type 2 diabetes mellitus without complications
Anemia, unspecified
Gastro-esophageal reflux disease without esophagitis
Fever, unspecified
Pneumonia, unspecified organism
Encounter for screening for other viral diseases
Acute kidney failure, unspecified
Personal history of nicotine dependence
Anxiety disorder, unspecified
Other nonspecific abnormal finding of lung field
Atherosclerotic heart disease of native coronary artery without angina pectoris
Obesity, unspecified
Chest pain, unspecified
Hypoxemia

Though the Havervall article cited 4 of their most common symptoms, only fatigue and dyspnea were in our diagnosis dataset and not in the top 20 diagnosis codes (29th to 41st and 58th to 29th, respectively).

Diagnosis codes of interest might be shortness of breath, acute respiratory failure with hypoxia, pneumonia, anxiety and hypoxemia.

Creating the Boolean Array of Diagnosis and Medication Codes

	patient_id	U07.1	E11.9	E78.5	K21.9	R05	Z79.899	I25.10	Z00.00	E03.9	...	M10.9	E53
0	00306db74924cc5d59790cebb4954205c28e65ce	0	0	0	0	0	0	0	0	0	...	0	
1	0096e6391a4b58cf889d269779dfabc73e1f677b	0	0	0	0	1	0	0	0	0	...	0	
2	00b045496bccef7d94a1299ca8e3ed8f8e8d0a7d	0	0	1	0	0	1	1	0	0	...	0	
3	01754d44d7d60e6708de27b04d34c98840c83134	0	0	0	0	0	0	0	0	0	...	0	
4	022ec067c4f60a17d1d36f36e33e4afd89919cb7	0	0	0	0	0	0	0	0	0	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
418	ff92bc798dfc72b79179e9b4bb99149ef3e1c6a7	0	0	0	0	0	0	0	0	0	...	0	
419	ffb072aa489485bcece5b3d5f934662ef426ce8a	0	1	1	0	0	0	0	0	1	...	1	
420	ffb10f1db4dd012eceebc43d1fc8a9d44d5ca382	1	0	0	0	0	0	0	0	0	...	0	
421	ffb770615ddcc1bad13498c713f69f26e6d56798	0	0	0	0	0	0	1	0	0	...	0	
422	fff84b953f0967207d3a6d86b823f553b35dd60d	0	0	0	0	0	1	0	0	0	...	0	



Getting Started – Organizing by PID

Used Clustering Paper as reference

> "Journal of Biomedical Informatics 102 (2020) 103360"

Wanted to create array of binary values representing top diagnosis codes

Started with all diagnosis/medication codes with PID listed out

Groupby(PID)

Journal of Biomedical Informatics 102 (2020) 103360

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

ELSEVIER

Clustering datasets with demographics and diagnosis codes

Haodi Zhong[a], Grigorios Loukides[a,*], Robert Gwadera[b]

[a] Department of Informatics, King's College London, London, UK
[b] School of Computer Science, Cardiff University, Cardiff, UK

ARTICLE INFO

Keywords:
Clustering
Demographics
Diagnosis codes
Pattern mining

ABSTRACT

Clustering data derived from Electronic Health Record (EHR) systems is important to discover relationships between the clinical profiles of patients and as a preprocessing step for analysis tasks, such as classification. However, the heterogeneity of these data makes the application of existing clustering methods difficult and calls for new clustering approaches. In this paper, we propose the first approach for clustering a dataset in which each record contains a patient's values in demographic attributes and their set of diagnosis codes. Our approach represents the dataset in a binary form in which the features are selected demographic values, as well as combinations (patterns) of frequent and correlated diagnosis codes. This representation enables measuring similarity between records using cosine similarity, an effective measure for binary-represented data, and finding compact, well-separated clusters through hierarchical clustering. Our experiments using two publicly available EHR datasets, comprised of over 26,000 and 52,000 records, demonstrate that our approach is able to construct clusters with correlated demographics and diagnosis codes, and that it is efficient and scalable.

	Unnamed: 0	patient_id	encounter_id	code_system	code	date	derived_by_TriNetX
0	1	c11adeff031d0701c258e12d950b9e090545875f	7689171769ec1a373b96cdfd751a6742aa940cc4	ICD-10-CM	R06.02	20200811	F
1	826	4aa5ad8abce18968c9416141e0c269c94749ce4e	8522b86068a3fd6bcc255f01ff9ec23411fc91ba	ICD-10-CM	A41.9	20200425	F
2	830	4aa5ad8abce18968c9416141e0c269c94749ce4e	8522b86068a3fd6bcc255f01ff9ec23411fc91ba	ICD-10-CM	D64.9	20200420	F
3	838	4aa5ad8abce18968c9416141e0c269c94749ce4e	5ec208e755a6b81cd1b197e3ae8b293762ec074e	ICD-10-CM	I50.9	20200427	F
4	859	4aa5ad8abce18968c9416141e0c269c94749ce4e	82c7f257b49f7cff06759881a6ca766c78bdd521	ICD-10-CM	J96.01	20200713	F
...	...	...	...	...	...	...	...
3795	9307708	662c6ced056c100dfd03ca5f00c6ee3690555886	9a8e0d2fae0382a29df9f84db1543131a19e4c9c	ICD-9-CM	V70.0	20200908	F
3796	9307709	662c6ced056c100dfd03ca5f00c6ee3690555886	ac8c0893a907cbb6b6cb8e555508cb0d603158e4	ICD-9-CM	V70.0	20200908	F
3797	9307710	662c6ced056c100dfd03ca5f00c6ee3690555886	22ade516fa8ddbceed44e1784e4e503f70ef12eb	ICD-9-CM	V70.0	20200929	F
3798	9307711	662c6ced056c100dfd03ca5f00c6ee3690555886	ceb03ca9101874e68dff6e553356d5b293f4fdff	ICD-9-CM	V70.0	20200929	F
3799	9307722	662c6ced056c100dfd03ca5f00c6ee3690555886	5385595c44710b4073ff3e546a89eb67d3f9ff07	ICD-9-CM	V76.12	20200908	F

Creating the Binary Values

Create dataframe header with PID and all diagnosis/medication codes

Loop through all patients:

If PID has code, add 1 in that Code column

Otherwise add a 0 in the column

Creates Dataframe w/ PID and all codes for each PID

Merge this dataframe with the Master Dataframe

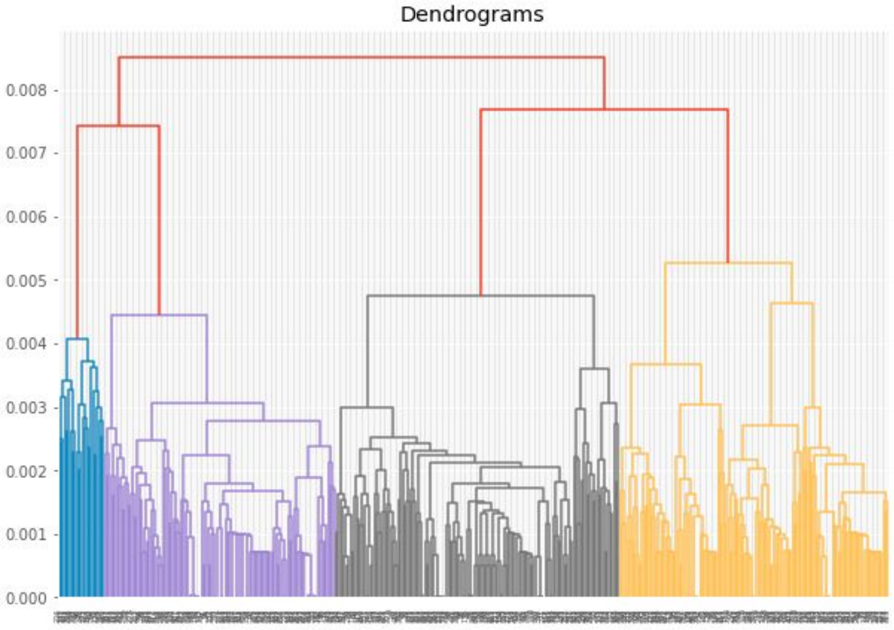
```
#create a function that returns PID and boolean array
def get_array(master, check, patient_ID):
    bool_list = [patient_ID]
    for i in master:
        if i in check:
            bool_list.append(1)
        else:
            bool_list.append(0)
    return bool_list

#function to create new row as df
def add_row(row, df):
    df_header = list(df.columns)
    row_to_add = pd.DataFrame([row], columns=df_header)
    return row_to_add
```

```
#loop through all pid and dx in list2 and create boolean array
for count, patient in enumerate(list2):
    print(str(count/(len(list2))) + 'precent done')
    master = res[0:21]
    check = list(list2[count][1])
    patient_ID = (list2[count][0])
    x = get_array(master, check, patient_ID)
    y = add_row(x, master_df)
    master_df = master_df.append(y, ignore_index=True)
```

Applications – Clustering

- With this cleaned and organized dataset, clustering exercise can be performed on the dataset
  - This could be used to group patients into clusters based on their diagnosis codes and demographic information
- Could be used to make informed decisions about what contributes to long COVID and what are its most common outcomes

```
import scipy.cluster.hierarchy as shc
plt.figure(figsize=(10, 7))
plt.title("Dendrograms")
dend = shc.dendrogram(shc.linkage(X_scaled, method='ward'))
```



Challenges and Solutions

- Size of Dataset files - Chunking
- SQLite Trial and Failure
- Time Management - We weren't aware of the amount of time needed to filter the dataset, our solution was to use 10% of the patients to prove our concept.

```
def row_count(input):
    with open(input) as f:
        for i, l in enumerate(f):
            pass
    return i
row_count('lab_result.csv')
```

```
630309878
```

```
In [10]: # start a timer
begin_time = datetime.datetime.now()

# append chunks to an empty list
chunks = []
for i in range(int(0.1 * len(PID))):
    chunks.append(df[(df['patient_id'] == PID[i] ) & (df['date']<dates[i])])

# concatenate chunks to df and filter for 200 most common codes
pre_covid = pd.concat(chunks)
pre_covid = pre_covid[pre_covid['code'].isin(code_list)]

# write to csv
pre_covid.to_csv('pre_covid.csv')

# end timer and print time
print(datetime.datetime.now() - begin_time)

12:38:48.011970
```