COHD-COVID: Publicly Shared COVID-19 Statistics Mined from EHR Data

Junghwan Lee, MA[1]‡, Casey Ta, PhD[1]‡, Cong Liu, PhD[1], Jae Hyun Kim, PhD[1], Chunhua Weng, PhD[1*]
[1]Columbia University, New York, NY

‡: Equal contribution *: Corresponding author

Abstract

Researchers and clinicians across the globe are dedicating enormous efforts towards the battle against coronavirus disease 2019 (COVID-19). Secondary analysis of clinical data can facilitate these research efforts since important patient characteristics can often be found by examining the shared clinical data. We leverage the previously developed Columbia Open Health Data (COHD) analysis pipeline and data infrastructure to present COHD-COVID, a publicly accessible application programming interface (API) that provides electronic health record (EHR) prevalence of COVID-19 patients derived from Columbia University Irving Medical Center. COHD-COVID provides prevalence rates of conditions and drugs from a cohort of COVID-19 hospitalized patients. The EHR prevalence rates are provided from the full cohort and further stratified by age (18-64 vs 65+) and sex. Here, we report some informative statistics mined from COHD-COVID, including the most common conditions and drugs observed in the COVID-19 patients cohort. Development is ongoing to provide comparative baseline prevalence rates, including hospitalized influenza patients. These data will be made publicly available via the COHD API for computational analyses and data exploration supporting COVID-19 research.

Research Category: Open-source analytics development

Background

The coronavirus disease 2019 (COVID-19) global pandemic has sparked massive research efforts in the fight against the novel disease, including characterizing the disease and clinical progression, identifying risk factors for hospitalization, and finding drugs that can be repurposed to lessen disease severity[1-3]. Utilization of clinical data from different institutions, hospitals, and nations can facilitate these research efforts since important characteristics of the patients are often found by examining the shared clinical data. Despite the immediate need, publicly accessible clinical data on COVID-19 remain limited in the United States[4].

Columbia Open Health Data (COHD) can accelerate translational research by providing open access to observational statistics about conditions, drugs, procedures, and demographics derived from electronic health records (EHR) from Columbia University Irving Medical Center (CUIMC)[5-7]. CUIMC serves the large and diverse population of New York City and its surrounding areas, which has been an epicenter of the disease since the first COVID-19 patient was confirmed on March 1, 2020. With the aim of providing sharable clinical data to catalyze future COVID-19 research, we present COHD-COVID, a publicly accessible API providing EHR prevalence of COVID-19 patients derived from CUIMC.

Methods

Data processing

CUIMC transformed its clinical data warehouse containing inpatient and outpatient data to OMOP CDM v5 in June 2020. We identified a cohort of hospitalized COVID-19 patients in CUIMC's OMOP database based on the OHDSI network study definitions[1]. For each cohort, we extracted condition, drug, and procedure concepts from the OMOP standardized clinical data tables and performed the following EHR prevalence analyses 1) on the full cohort, 2) stratified by sex, and 3) stratified by age (18-64 vs. 65+). The concept prevalence, P^C, was defined as Eq(1):

$$P^C = \frac{|T_C|}{|T_H|} \quad Eq(1)$$

where T_C was the set of unique patients observed with concept C, and T_H was the set of unique patients in the cohort *H.* We also calculated hierarchical counts by defining T_C as the set of unique patients observed with concept C or any its descendants defined in the `concept_ancestor` table. For protection of patient privacy, we excluded rare concepts observed in 10 or fewer patients and perturbed the exact counts using Poisson randomization[5].

Results

Here, we report some informative statistics mined from COHD-COVID. **Figure 1** shows the top 10 condition concepts with the highest prevalence in the COVID-19 cohort. Prevalence rates for this figure were calculated without hierarchical analysis (e.g. prevalence of Dyspnea does not incorporate descendant concepts) and are shown among the full cohort as well as stratified by age and sex. **Figure 2** shows the top 10 drug ingredients with the highest prevalence rates in the COVID-19 cohort. For drug concept prevalence, we report the prevalence of RxNorm ingredients considering hierarchical relationships between concepts to aggregate exposures of the same drug ingredient from different dosages and formulations. Drug exposure prevalence rates are also stratified by age and sex.

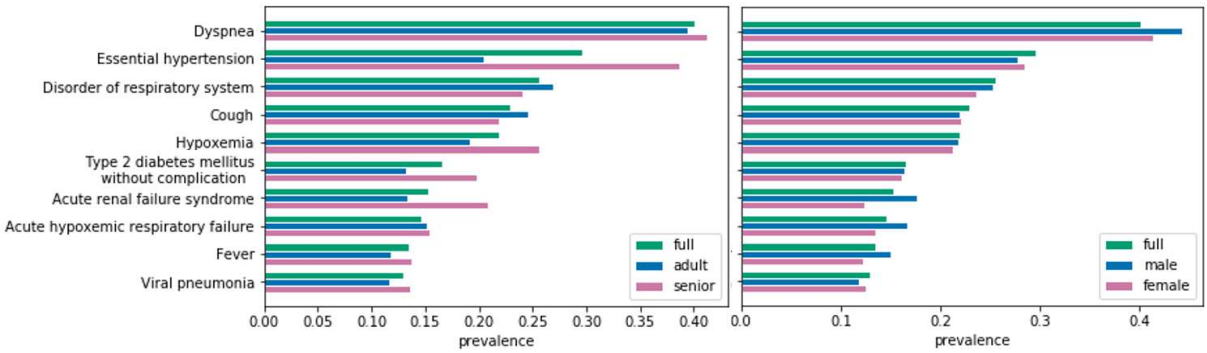

Figure 1. Condition concept prevalence in (left) age (18-64 vs 65+) and (right) sex subgroups of the COVID-19 patient cohort.

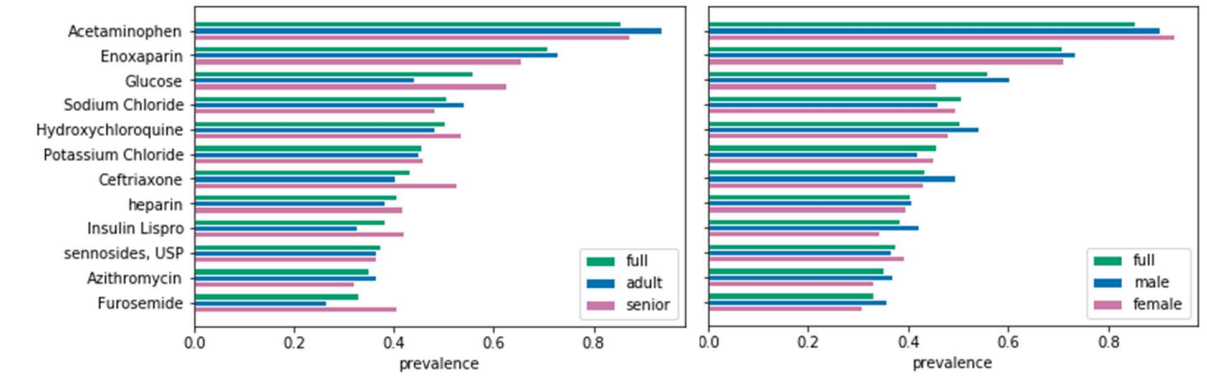

Figure 2. Drug concept prevalence in (left) age (18-64 vs 65+) and (right) sex subgroups of the COVID-19 patient cohort.

Discussion and Conclusions

The conditions and drugs commonly observed among hospitalized COVID-19 patients are in accordance with published findings[1-2], including hypertension and diabetes mellitus, which may increase the risk of developing COVID-19 infection[8], and acute kidney injury, which may be caused by infection of renal podocyte and proximal straight tubule cells[9,10]. Although most conditions were observed at higher rates in senior patients relative to adult patients, disorder of respiratory system and cough were recorded at higher rates in adults. Most of the common conditions were observed evenly between both sexes with the exception of acute renal failure syndrome, which is elevated in male patients in the general cohort[5]. Acetaminophen was the most common drug exposure in COVID-19 patients. Acetaminophen is an antipyretic agent that has been recommended over ibuprofen for reducing fever in COVID-19 patients due to potential safety concerns with ibuprofen[11]. Enoxaparin, an anticoagulant, was the second most common drug, as coagulopathy is an important complication of COVID-19[12].

Development of the COHD-COVID resource is ongoing. In addition to the EHR prevalence rates provided for the COVID-19 cohort, we plan to generate comparative data for two additional baseline cohorts: and general patients and patients hospitalized with influenza. We also plan to calculate concept-pair co-occurrences and their association metrics within each of these three cohorts to identify, for example, comorbid conditions within the COVID-19 cohort associated with ventilation procedures or death. COHD-COVID will be accessible via a web API to facilitate computational analyses and data exploration supporting COVID-19 research. As a data source of clinical observations derived from the largest hospital in New York City, we hope COHD-COVID will provide researchers and clinicians fighting against the pandemic with valuable resources.

Acknowledgement

The study was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number OT2TR003434. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Burn E, You SC, Sena A, et al. An international characterisation of patients hospitalised with COVID-19 and a comparison with those previously hospitalised with influenza. *medRxiv.* 2020.
2. Zhou F, Yu T, Du R, et al. Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study. *The lancet.* 2020.
3. Shah B, Modi P, Sagar SR. In silico studies on therapeutic agents for COVID-19: Drug repurposing approach. *Life Sciences.* 2020:117652.
4. Argenziano MG, Bruce SL, Slater CL, et al. Characterization and clinical course of 1000 patients with coronavirus disease 2019 in New York: retrospective case series. *bmj.* 2020;369.
5. Ta CN, Dumontier M, Hripcsak G, Tatonetti NP, Weng C. Columbia Open Health Data, clinical concept prevalence and co-occurrence from electronic health records. *Scientific data.* 2018;5:180273.
6. Ahalt SC, Chute CG, Fecho K, et al. Clinical data: sources and types, regulatory constraints, applications. *Clinical and translational science.* 2019;12(4):329.
7. Fecho K, Ahalt SC, Arunachalam S, et al. Sex, obesity, diabetes, and exposure to particulate matter among patients with severe asthma: Scientific insights from a comparative analysis of open clinical data sources during a five-day hackathon. *Journal of biomedical informatics.* 2019;100:103325.
8. Fang L, Karakiulakis G, Roth M. Are patients with hypertension and diabetes mellitus at increased risk for COVID-19 infection? *The Lancet Respiratory Medicine.* 2020;8(4):e21.
9. Hirsch JS, Ng JH, Ross DW, et al. Acute kidney injury in patients hospitalized with COVID-19. *Kidney International.* 2020.
10. Pan X-w, Da Xu HZ, Zhou W, Wang L-h, Cui X-g. Identification of a potential mechanism of acute kidney injury during the COVID-19 outbreak: a study based on single-cell transcriptome analysis. *Intensive care medicine.* 2020:1.
11. Sodhi M, Etminan M. Safety of ibuprofen in patients with COVID-19: causal or confounded? *Chest.* 2020.
12. Fogarty H, Townsend L, Ni Cheallaigh C, et al. COVID19 coagulopathy in Caucasian patients. *British journal of haematology.* 2020.