RESEARCH PROTOCOL:

Characterising and comparing individuals with COVID-19 or influenza: protocol for an OHDSI network study

Table of Contents

[1. List of Abbreviations 2](#_heading=h.17dp8vu)

[2. Responsible Parties 2](#_heading=h.3rdcrjn)

[2.1. Investigators and Authors 2](#_heading=h.26in1rg)

[2.2 Sponsor 2](#_heading=h.lnxbz9)

[3. Abstract 3](#_heading=h.35nkun2)

[4. Amendments and Updates 3](#_heading=h.1ksv4uv)

[5. Rationale and Background 3](#_heading=h.1fob9te)

[6. Objective 3](#_heading=h.3znysh7)

[7. Methods](#_heading=h.2et92p0) 4

[7.1 Data Sources](#_heading=h.2jxsxqh) 4

[7.2 Study design 5](#_heading=h.tyjcwt)

[7.3 Target cohorts 5](#_heading=h.3dy6vkm)

[7.4 Stratifications](#_heading=h.4d34og8) 6

[7.5 Features of interest](#_heading=h.1y810tw) 8

[7.6 Logistics of Executing a Federated Analysis](#_heading=h.4i7ojhp) 8

[8. Sample Size and Study Power](#_heading=h.2xcytpi) 9

[9. Strengths and Limitations](#_heading=h.1ci93xb) 9

[9.1 Strengths](#_heading=h.3whwml4) 9

[9.2 Limitations](#_heading=h.2bn6wsx) 9

[10. Protection of Human Subjects](#_heading=h.qsh70q) 9

[11. Management and Reporting of Adverse Events and Adverse Reactions](#_heading=h.3as4poj) 9

[12. Plans for Disseminating and Communicating Study Results](#_heading=h.1pxezwc) 9

[13. Example tables 10](#_heading=h.49x2ik5)

[Table 1: Demographics](#_heading=h.147n2zr) 10

[Table 2: Conditions](#_heading=h.ihv636) 10

[Table 3: Medications](#_heading=h.41mghml) 10

[References](#_heading=h.2grqrue) 12

[Appendix 1: Cohort Definitions 1](#_heading=h.vx1227)3

# 1. List of Abbreviations

|  |  |
| --- | --- |
| AKI | Acute Kidney Injury |
| ARDS | Acute Respiratory Distress Syndrome |
| CDM | Common Data Model |
| COVID-19 | 2019 novel coronavirus / coronavirus disease 2029 |
| DVT | Deep Venous Thrombosis |
| EHDEN | European Health Data and Evidence Network |
| HIV | Human Immunodeficiency Virus |
| OMOP | Observational Medical Outcomes Partnership |
| OHDSI | Observational Health Data Science and Informatics |
| PE | Pulmonary Embolism |
| RxNorm | US-specific terminology in medicine that contains all medications available on the US market |
| SARS-CoV-2 | Severe Acute Respiratory Syndrome-Coronavirus-2 |
| SNOMED | Systematized Nomenclature of Medicine |
| WHO | World Health Organization |

# 

# 2. Responsible Parties

## 2.1. Investigators and Authors

|  |  |
| --- | --- |
| **Investigator/Author** | **Institution/Affiliation** |
| Talita Duarte-Salles\* | Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain |
| Albert Prats-Uribe\* | Centre for Statistics in Medicine, NDORMS, University of Oxford |
| Daniel Prieto-Alhambra | 1) Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), 2) Centre for Statistics in Medicine, NDORMS, University of Oxford |
| Seng Chan You | Department of Biomedical Informatics, Ajou University |
| George Hripcsak | Department of Biomedical Informatics, Columbia University |
| Patrick B. Ryan | 1) Janssen Research and Development, Titusville, NJ, USA, 2) Columbia University, New York, NY, USA |
| Kristin Kostka\* | Real World Solutions, IQVIA |
| Anthony G. Sena | 1) Janssen Research and Development, Titusville, NJ, USA, 2) Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands |
| Edward Burn | IDIAPJGol, Barcelona, Spain; Universitat Autònoma de Barcelona, Spain |
| Gowtham Rao | Janssen Research and Development, Titusville, NJ, USA |
| Azza Shoaibi | Janssen Research and Development, Titusville, NJ, USA |

\* Co-Principal investigator

## 

Authorship will vary from paper to paper. The Responsible Parties involved in this protocol take accountability for the overarching protocol, package development, providing assistance to sites running the analysis and ensuring site-specific governance is adhered to in all publications generated from this protocol.

For subgroup analysis, the following table represents individuals who will be contributing to follow-on work to write manuscripts on these specific topics. This list is not exhaustive of all possible authors for future papers:

|  |  |  |
| --- | --- | --- |
| **Investigator/Author** | **Institution/Affiliation** | **Topics** |
| David Vizcaya | Bayer Pharmaceuticals, Barcelona, Spain | COVID-19 in paediatric population  COVID-19 in CKD patients |
| Daniel R. Morales | Division of Population Health and Genomics, University of Dundee, UK | COVID-19 in community patients  COVID-19 in asthma/COPD |
| Asieh Golozar | Regeneron Pharmaceuticals, NY USA; Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA | COVID-19 testing |
| Andrea V Margulis | RTI Health Solutions | COVID-19 in pregnancy |
| Julianna Kohler | United States Agency for International Development, Washington, DC, USA | COVID-19 in HIV/Immunocompromised (inclusive of Hepatitis C) and tuberculosis populations |
| Talita Duarte-Salles | Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain | COVID-19 in cancer patients |
| Edward Burn | Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain; Universitat Autònoma de Barcelona, Spain | COVID-19 in hospitalised adults  COVID-19 and intensive services (ECMO, ventilation, tracheostomy) |
| Seng Chan You | Department of Biomedical Informatics, Ajou University | COVID-19 uncomplicated hypertension |
| Martina Recalde | Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain; Universitat Autònoma de Barcelona, Spain | COVID-19 in hypertension, type 2 diabetes patients, obesity |
| Jeroen de Bont | Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain; ISGlobal, Barcelona Institute for Global Health, Spain | COVID-19 in hypertension, type 2 diabetes patients, obesity |
| Maura Beaton | Columbia University, New York, NY USA | COVID-19 gender differences |
| Noemie Elhdad | Columbia University, New York, NY USA | COVID-19 gender differences |
| Kristin Kostka | Real World Solutions, IQVIA, Cambridge, MA | COVID-19 and thrombolytic events  COVID-19 and cardiac outcomes |
| Vojtech Huser | National Institutes of Health, Bethesda, MD | COVID-19 follow-up time |

## 2.2 Sponsor

This study was undertaken by Observational Health Data Science and Informatics (OHDSI), an open collaboration. **DPA** receives partial support from the National Institute for Health Research Oxford Biomedical Research Centre and Senior Research Fellowship. **SCY** receives grant funding from the Korean Ministry of Health & Welfare and from the Korean Ministry of Trade, Industry & Energy. **GH** receives grant funding from the US National Institutes of Health and contracts from Janssen Research and Development. **PBR**, **AGS, GR** and **AS** are employees of Janssen Research and Development and shareholders in Johnson & Johnson. **KK** is an employee of IQVIA. **APU** receives grants from MRC and Fundacion Alfonso Martin Escudero. Participating data owners will be responsible for self-reporting any grants funding the conversion and maintenance of their OMOP CDM. Disclosures will be reported in accordance with publication policies of journals papers are submitted to. No other disclosures reported.

# 3. Abstract

In this study we will describe the baseline demographic and clinical characteristics, as well as the occurrence of treatments and outcomes of individuals tested or with COVID-19 in 2019-2020. We will compare these with the characteristics of individuals with Influenza in 2017-2018.

# 4. Amendments and Updates

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Number** | **Date** | **Section of study protocol** | **Amendment or update** | **Reason** |
| V2 | 27APR2020 | All | Incorporating community feedback | Collaboration |
| V3 | 04MAY2020 | All | Incorporating final changes from core team | Consistency |

# 5. Rationale and Background

Since December 31st 2019, when the first case was reported to WHO, infections by SARS-CoV-2 have spread worldwide. Our current understanding of the disease's clinical outcomes is limited to case series in hospitalised populations. The baseline characteristics of 138 patients with Coronavirus disease 2019 (COVID-19) hospitalized in Wuhan due to pneumonia, for example, have been reported.1 In another study, a series of 21 cases admitted to an ICU in Washington state reported clinical and demographic characteristics at hospital admission.2 A more recent study has described the baseline characteristics and complications in 393 confirmed COVID-19 cases admitted to two hospitals in New York City. 3

Government and healthcare strategies for testing and diagnosing suspected COVID-19 patients vary extensively between countries and by time. Some countries have been urged to implement widespread testing following WHO's recommendation, and other countries are testing fewer cases often due to lack of availability of laboratory tests.4 Often these strategies focus on testing more severe and high-risk patients. Selective testing might lead to overestimation of poor outcomes and mortality. An understanding of the characteristics of tested populations by country government policies would better inform epidemiology studies and understanding of disease burden, magnitude and outcomes.

As more routinely-collected data emerges on individuals with COVID-19, a federated network study could provide a fuller picture of these individuals, particularly among specific population groups that might be more susceptible to COVID-19 and for which limited information is currently available, such as the paediatric or pregnant populations, or persons with comorbidities such as diabetes, hypertension or cancer. Moreover, the use of such data to inform comparisons with previous viral outbreaks, like regular seasonal influenza epidemics, could also help us to better understand the profiles of individuals with COVID-19. Data assets mapped to the Observational Medical Outcomes Partnership (OMOP) common data model (CDM)5 provide a unique opportunity to make a difference in the current crisis, allowing for robust analyses to be performed in a timely across a network of sites.

# 6. Objective

The primary objective of this study is to describe the baseline demographic and clinical characteristics, as well as treatments and occurrence of outcomes of interest among individuals tested for SARS-CoV-2 or diagnosed with COVID-19 after December 1st 2019, overall and by sex, age and comorbidities. We will also describe the characteristics and outcomes of the population with seasonal influenza infection between September 1st 2016 and April 1st 2018 as a benchmark.

(***Note:*** *For computational efficiency, this analysis will deviate from the prior OHDSI Hospitalized in COVID-19 analysis which included a larger sample of years (2014-2019) of the seasonal influenza population and a separate cohort for H1N1 (2009-2010). The computational run time to pull these cohorts required a significant investment by data owners and, at the time of authoring this analysis, it was decided a smaller window of years – particularly years in which flu severity was highest – would be the most efficient use of compute.)*

# 7. Methods

### 7.1 Data Sources

This study is a multinational cohort study describing the baseline characteristics, treatments and outcomes of 12 cohorts, namely:

1. SARS-CoV-2 tested
2. SARS-CoV-2 tested positive
3. SARS-CoV-2 tested with a COVID-19 diagnosis record or a positive test
4. COVID-19 diagnosed or with a positive test
5. COVID-19 diagnosed/positive tested and hospitalised
6. COVID-19 diagnosed/positive tested, hospitalised, and requiring intensive services
7. Influenza between 2017-2018 diagnosed or with a positive test
8. Influenza between 2017-2018 diagnosed/positive tested and hospitalized
9. Influenza between 2017-2018 diagnosed/positive tested, hospitalised, and requiring intensive services

A South Korean national claims database, multiple U.S. (New York City, San Francisco, Boston, Denver, US Veterans’ Affairs) health systems databases, and one Spanish EHR database are accumulating COVID-19 patients and have tested the operability of our analysis package at their sites (**Table 1**). As more data becomes available, we will include additional databases that are formatted to the OMOP-CDM. These will reflect the clinical experience of patients from seven European countries (Belgium [general practice EHR], Netherlands [general practice EHR], Germany [general practice EHR, hospital EHR], France [general practice EHR, outpatient specialist EHR], Spain [general practice EHR, outpatient specialist EHR and linked inpatient administrative data], Estonia [EHR, claims, and registry data], and the United Kingdom [general practice EHR, linked to hospital EHR]), the United States of America (general practice EHR, outpatient specialist EHR, hospital EHR, insurance claims], South Korea [EHR, claims, and registry data], Japan [insurance claims], and any other country with mapped data willing to participate.

The study will be conducted using data from real world data sources that have been mapped to the OMOP Common Data Model in collaboration with the Observational Health Data Sciences and Informatics (OHDSI) and European Health Data and Evidence Network (EHDEN) initiatives. The OMOP Common Data Model (<https://github.com/OHDSI/CommonDataModel/wiki>) includes a standard representation of health care experiences (such as information related to drug utilization and condition occurrence), as well as common vocabularies for coding clinical concepts, and enables consistent application of analyses across multiple disparate data sources.6

**Table 1**. Data sources formatted to the OMOP CDM that as of 20 Apr 2020 include COVID-19 patients.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Data source** | **Source population** | **Sample size** | **Data type** | **Longitudinal history** |
| South Korea: Health Insurance and Review Assessment (HIRA) | All citizens in South Korea | ≈ 50 million | Administrative fee-for-service claims data collected for healthcare reimbursement, including healthcare services such as treatments, pharmaceuticals, procedures, and diagnoses. | 5-years of available look-back (data older than 5-years is deleted from the database) |
| Columbia University Irving Medical Center | Patients of the Columbia University Irving Medical Center (New York City, USA) | ≈ 6 million | General practice electronic health records, outpatient specialist electronic health records, inpatient hospital electronic health records, hospital billing/summary | 1989 (1978 for diagnoses) |
| Stanford Medicine Research Data Repository (STARR) | Patients of Stanford University (San Francisco Bay Area, USA) | ≈ 3 million | General practice electronic health records, outpatient specialist electronic health records, inpatient hospital electronic health records, hospital billing/summary | 2008 |
| Tufts Clinical Academic Research Enterprise Trust (CLARET) | Patients of Tufts Medical Center (Boston, USA) | ≈ 1 million | Electronic medical record data on approximately 1 million patients who received care beginning in 2006 at Tufts Medical Center (TMC). TMC is an academic medical center that includes Tuft Medical Center's main downtown Boston hospital for adult patients, the Floating Hospital for Children, and associated primary and specialty care clinics. CLARET contains TMC's EHR data fused with data on the same patients from TMC's CoC accredited tumor registry, its oncology EHR, and death data from the Massachusetts State Registry of Vital Statistics. EHR data streams ingested into CLARET include controlled vocabulary data on all domains except cost, and select free text sources and devices. | 2006 |
| Veteran’s Affairs | Patients of the Veterans Affairs in the United States | ≈ 9 million | VA OMOP data reflects the national Department of Veterans Affairs health care system, which is the largest integrated provider of medical and mental health services in the United States. Care is provided at 170 VA Medical Centers and 1,063 outpatient sites serving more than 9 million enrolled Veterans each year. | To be confirmed |
| UC Health Data Compass | Patients of UC Denver and Children’s Hospital of Colorado (Denver, USA) | To be confirmed | Health Data Compass (HDC) is a multi-institutional data warehouse. HDC contains inpatient and outpatient electronic medical data including patient, encounter, diagnosis, procedures, medications, laboratory results from two electronic medical record systems (UCHealth and Children's Hospital of Colorado), state-level all-payers claims data, and the Colorado death registry. | To be confirmed |
| University of Washington | Patients of UW Medical Center (Seattle, USA) | ≈ 5 million | The clinical data warehouse of University of Washington Medical Center - Montlake, Harborview Medical Center, UW Medical Center – Northwest, Seattle WA, based on its current electronic health record systems, with data spanning over 10 years and including roughly 5 million patients. We also have de-identified and annotated xrays for COVID-19 patients who were imaged in a specific COVID-19 PACS. | 2009 |
| The Information System for Research in Primary Care (SIDIAP) | General population in Catalonia, Spain | ≈ 8 million | Electronic health records from primary care partially linked to inpatient data. SIDIAP is also linked to pharmacy dispensations and primary care laboratories. Healthcare is universal and tax-payer funded in the region, and primary care physicians are gatekeepers for all care and responsible for repeat prescriptions. | 2006 |

### 7.2 Study design

The study will be an observational cohort study based on routinely-collected health care data which has been mapped to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM). Cohorts of individuals with COVID-19 or influenza will be identified. Characteristics of these individuals at their index date will be described. Treatments and outcomes of these individuals after their index date will be described.

### 7.3 Target cohorts

COVID-19 cohorts

**Target Cohort #1:** participants in the *tested SARS-CoV-2 cohort* will:

* have a record of a first test for SARS-CoV-2 (index event) after December 1st 2019.

**Target Cohort #2:** participants in the *tested positive SARS-CoV-2 cohort* will:

* have a record of a positive test for SARS-CoV-2 (index event will be the earliest test date that occurs within 7 days of positive test result) after December 1st 2019.

***Note: For computational efficiency, there will not be a separate tested negative SARS-CoV-2 cohort.***

**Target Cohort #3:** participants in the *tested SARS-CoV-2 with a COVID-19 diagnosis record or a positive test result cohort* will:

* have a record of a test for SARS-CoV-2 (index event first test) after December 1st 2019, and either,
  + have a record of COVID-19 diagnosis, or
  + have a record of a positive test for SARS-CoV-2.

**Target Cohort #4:** participants in the *COVID-19 diagnosed or with a positive test cohort* will:

* have a record of a positive test for SARS-CoV-2 after December 1st 2019, or,
* have a record of a COVID-19 diagnosis (index date is date of the first of these 2 events).

**Target Cohort #5:** participants in the *hospitalised with COVID-19 cohort* will:

● have a hospitalisation (index event) after December 1st 2019, and,

● a record of a COVID-19 diagnosis or a SARS-CoV-2 positive test result in the 3 weeks prior and up to end of hospitalisation, and,

● have no COVID-19 associated hospitalisation in the six months prior to the index event.

**Target Cohort #6:** participants in the *hospitalised with COVID-19 and requiring intensive services cohort* will:

● have a hospitalisation (index event) after December 1st 2019, and,

● a record of Ventilation or Tracheostomy or ECMO during hospitalization, and,

● a record of a COVID-19 diagnosis or a SARS-CoV-2 positive test result in the 3 weeks prior and up to end of hospitalisation, and,

* have no COVID-19 associated hospitalisation in the six months prior to the index event.

Influenza cohorts

**Target Cohort #7:** participants in the *influenza diagnosed or with a positive test cohort* will:

* have a record of a positive test for influenza between September 1st 2017 and April 1st 2018, or,
* have a record of influenza diagnosis (index on first event).

**Target Cohort #8:** participants in the *hospitalised with influenza cohort* will:

● have a hospitalisation (index event) between September 1st 2017 and April 1st 2018,

● with a record of a diagnosis/positive test of influenza in the 3 weeks prior and up to end of hospitalisation,

● have no influenza associated hospitalisation in the six months prior to the index event.

**Target Cohort #9:** participants in the *hospitalised with influenza and require intensive services cohort* will:

● have a hospitalisation (index event) between September 1st 2017 and April 1st 2018, and

● a record of Ventilation or Tracheostomy or ECMO during hospitalization, and,

● have a record of a diagnosis/positive test of influenza in the 3 weeks prior and up to end of hospitalisation, and,

● have no influenza associated hospitalisation in the six months prior to the index event.

These cohorts will all be identified without any requirement for prior observation time, and also with the added restriction of having a minimum of 365 days of prior observation time available in order to assess how large proportion of subjects fall ill without previous healthcare encounters, and if their inclusion in analyses would change results or conclusions.

### 7.4 Follow-up

Different index dates will be used for each of the cohorts above:

**Target Cohort #1:**

Index event: First test for COVID-19/influenza.

Follow-up: 30 days after the index event.

**Target Cohort #2:**

Index event: Earliest test date that occurs within 7 days of positive test result.

Follow-up: 30 days after the index event.

**Target Cohort #3:**

Index event: Earliest test date that occurs within 7 days of positive test result or COVID-19/influenza diagnosis.

Follow-up: 30 days after the index event.

**Target Cohort #4,7:**

Index event: First positive test result for COVID-19/Flu or first COVID-19/Flu Diagnosis (whichever appears first).

Follow-up: 30 days after the index event.

**Target Cohort #5,8:**

Index event: First hospitalisation, with a record of COVID-19/Flu 3w prior.

Follow-up: 30 days after the index event.

**Target Cohort #6,9:**

Index event: First hospitalisation, with a record of COVID-19/Flu 3w prior, and Ventilation or Tracheostomy or ECMO.

Follow-up: 30 days after the index event.

### 7.5 Stratifications

Each target cohort will be analysed in full and stratified on factors based on the following pre-index characteristics, all stratum are pending meeting minimum reportable cell counts (as specified by data owners):

* Follow-up time: overall, with full 30 days follow-up, without full 30 days follow-up
* Sex (Male vs. Female)
* All reportable age groups as well as specifically:
  + Pediatrics (Age <18). *If sample size allows, results will be reported stratified in the following age categories: overall (0 up to 17 years); neonates (0-28d), infants (29d- less than 1 year), toddler (1-less than 2y), early childhood (2-5y), middle childhood (6-11y), and adolescents (12-17y).* In case of limited sample size for granular classification, the following categories will be used: neonates, infants, and children (0-12y); and adolescents (12-17y)
  + Elderly (Age >= 65). If sample size allows, results will be reported stratified in the following age categories: 65-84 years, and >=85 years, or in finer age strata (65-69, 70-74, 75-79, 80-84, >=85 years)
* Type 2 Diabetes Mellitus
* Hypertension
* Chronic obstructive pulmonary disease (COPD)
* Asthma
* Chronic kidney disease
* End-stage renal disease
* Heart disease
* Cancer
* Pregnant women
* Human immunodeficiency virus infection
* Tuberculosis
* Hepatitis C
* Overweight/obese
* Smoker
* Disorder of nervous system
* Mental disorder
* Dementia
* Have Cardiovascular OR metabolic OR malignant history

### 7.6 Features of interest

These features span across the full set of target cohorts and research questions of interest in subgroups (incl. pediatrics, pregnant women, etc) so some features will only be relevant in some target cohorts or some subgroups, but the full list is given here.

Pre-index characteristics

These features will be described as assessed in two different time windows: the last 30 days (-1 to -30 days) and the year (-1 to -365 days) pre-index:

**Demographics**:

- Age: calculated as year of cohort start date – year of birth and with 5 year groupings

- Sex

**Concept-based:**

- Condition groups (SNOMED + descendants), >=1 occurrence during the interval

- Drug era groups (ATC/RxNorm + descendants), >=1 day during the interval which overlaps with at least 1 drug era

**Cohort-based: (to be added if relevant from lists)**

* Gestational diabetes
* Preeclampsia or eclampsia

Post-index characteristics

These features will be described in two different time windows: at index date (day 0) and in the 30 days from index date (0 to 30 days). As time elapses, additional windows of time will be investigated (eg. in the 60 days from index date [0 to 60 days]). The characteristics will include:

**Concept-based:**

* Condition groups (SNOMED + descendants), >=1 occurrence during the interval
* Drug era start groups (ATC/RxNorm + descendants), >=1 drug era start during the interval

**Cohort-based:**

General

* Initiation of intensive services during hospitalization
* Discharge from hospitalization

Symptoms

* Fever
* Cough
* Myalgia
* Malaise or fatigue
* Dyspnea
* Anosmia
* Hyposmia
* Dysgeusia

Outcomes by Condition

* Additional testing for COVID-19
* Hospitalisation
* Hospitalisation with pneumonia
* Hospitalisation with acute respiratory distress syndrome (ARDS)
* Hospitalisation with acute kidney injury (AKI)
* Hospitalisation with sepsis
* Hospitalisation with Venous thromboembolic (pulmonary embolism and deep vein thrombosis)
* Hospitalisation with heart failure
* Hospitalisation with arrhythmia (heart block/bradycardia, supraventricular tachycardia, ventricular tachycardia/ventricular fibrillation/cardiac arrest/sudden cardiac death)
* Hospitalisation with heart block/bradycardia
* Hospitalisation with supraventricular tachycardia
* Hospitalisation with ventricular tachycardia/ventricular fibrillation/cardiac arrest/sudden cardiac death
* Death
* Stillbirth
* Live birth: preterm delivery (< 37 weeks of gestational age)
* Live birth: postterm delivery (> 41 weeks of gestational age)
* Live birth: term delivery or unspecified delivery
* Abortion
* Premature preterm rupture of membranes
* Intrauterine growth restriction
* Hospitalisation for asthma exacerbation
* Hospitalisation for COPD exacerbation
* Pneumonia
* Chest pain or angina
* Angina
* Venous thromboembolic (pulmonary embolism and deep vein thrombosis) events
* Acute renal failure
* Hepatic failure
* Acute pancreatitis
* Heart failure
* Total cardiovascular disease events (ischemic stroke, hemorrhagic stroke, heart failure, acute myocardial infarction or sudden cardiac death)
* Cardiac arrhythmia
* Bradycardia
* Gastrointestinal bleeding
* Cardiovascular-related mortality
* Transient ischemic attack
* Stroke (ischemic or hemorrhagic)
* Acute myocardial infarction

Procedural treatments

* Hospitalisation requiring intensive services (ventilation OR tracheostomy OR ECMO)
* Hospitalisation requiring mechanical ventilation
* Hospitalisation requiring tracheostomy
* Hospitalisation requiring ECMO
* Hospitalisation requiring haemodialysis
* Hospitalisation requiring oxygen therapy
* Labor induction
* Cesarean section
* Emergency cesarean section

Pharmacological treatments (used during interval and no use in the 6 months prior to the interval)

* Hydroxychloroquine
* Hydroxychloroquine + azithromycin
* Hydroxychloroquine + fluoroquinolones
* Hydroxychloroquine + amoxicillin
* Hydroxychloroquine + ceftriaxone
* IL6 inhibitors (tocilizumab, sarilumab, siltuximab)
* tocilizumab
* Siltuximab
* sarilumab
* JAK inhibitors (baricitinib, tofacitinib)
* baricitinib
* tofacitinib
* TNF inhibitors
* adalimumab
* etanercept
* infliximab
* IL11 inhibitors (ustekinumab)
* IL1 inhibitor (Anakinra)
* HIV Protease inhibitors: lopinavir/ritonavir
* Interferon beta
* Interferon alpha
* Favipiravir
* Remdesivir
* Ivermectin
* warfarin, acenocoumarol
* heparins, enoxaparin
* DOAC (apixaban, rivaroxaban, dabigatran)
* Antithrombotics (clopidogrel, acetylsalicylic acid, ticlopidine, prasugrel, ticagrelor, triflusal, cangrelor, cilostazol, tissue plasmanegin activitor)
* Antifibrinolytics (tranexamic acid, aminocaproic acid)
* Ruxolitinib
* Siltuximab
* ACE inhibitors
* ARBs
* ndCCB
* thiazide diuretics
* Corticosteroids
  + Bevacizumab
  + Fingolimod
  + eculizumab
* Statins (HmgCOA reductase inhibitors)
* metformin
* DPP-4 inhibitors
* SGLT2 inhibitors
* GLP1 inhibitors
* H2 blockers (ranitidine, fomatadine)
* Alpha-1 blockers (Doxazosin, prazosin, terazosin, tasulosin, silodosin, alfuzosin)
* Immunoglobulins

### 7.7 Analysis: Characterizing cohorts

All analyses will be performed using the OHDSI Methods library. The code for this study can be found at [https://github.com/ohdsi-studies/Covid19MasterCharacterization](https://github.com/ohdsi-studies/Covid19HospitalizationCharacterization). Baseline covariates will be extracted using the FeatureExtraction package (<http://ohdsi.github.io/FeatureExtraction/>) to quantify Demographics (Gender, Prior Observation Time, Age Group), Condition Group Eras and Drug Group Eras (at and within 30 days after index date, at index date, within 30 days before index date, and within 365 days before index date). Additional cohort-specific covariates will be constructed using OMOP standard vocabulary concepts.

At the time of executing Feature Extraction, the package will create a data frame in which individuals’ age and sex will be extracted. Individuals’ medical conditions, procedures, measurements and medications will be summarized 1) over the year prior to their index date (-365d to -1d), 2) over the 30 days before index date (-30d to -1d), 3) at index date (0d), and 4) at and over the 30 days after index date (0d to 30d). Number and proportion of persons with feature variables during time-at-risk windows will be reported by target cohort and specific stratifications. Standardized mean differences (SMD) will be calculated when comparing characteristics of study cohorts, with plots comparing the mean values of characteristics for each of the characteristics (with the colour indicating the absolute value of the standardized difference of the mean).

### 7.8 Logistics of Executing a Federated Analysis

Sites will run the study analysis package locally on their data coded according to OMOP CDM. Only aggregate results will be shared with the study coordinator. Result files will be automatically staged into a ZIP file that can be transmitted using the OhdsiSharing R Library (<http://ohdsi.github.io/OhdsiSharing/>) or through a site’s preferred SFTP client using a site-specific key provisioned by the OHDSI Study Coordinator. Local data stewards are encouraged to review study parameters to ensure minCellCount function follows local governance. At a minimum, it is encouraged to keep this value to >5 to avoid any potential issues with re-identification of patients. An example of tables and figures can be seen in Section 13. (Note: covariates are constructed using controlled ontologies from the OMOP standard vocabularies though some labels may be replaced with publication-friendly labels due to space restrictions of the submitting journal.)

# 8. Sample Size and Study Power

No sample size or study power calculation are necessary for descriptive statistics.

# 9. Strengths and Limitations

### 9.1 Strengths

To our knowledge, this is one of the world’s largest observational sets of analyses of secondary health data for SARS-Cov-2 tested population and COVID-19 cases. We are running a multi-country, multi-center characterization study to understand baseline covariates, treatments and outcomes observed in COVID-19 and influenza cases. The use of a common data model and standard vocabularies ensures interoperability and portability of phenotypes utilized in this analysis. This will be the first time any analysis compares patient samples with these infectious conditions within the same population. The use of a federated study model will ensure no movement of patient-level data from institutions participating in this analysis. This is critically important to ensure the protection of patient privacy in the secondary use of routinely collected patient data. Data custodians will remain in control of the analysis run on these data and will conduct their own site-based validation processes to evaluate case reports against public health reporting.

### 9.2 Limitations

The potential for case misclassification of Influenza and COVID-19 is a concern, especially for COVID-19 for which testing plays a large part. COVID-19 cases may be underreported due to shortages in testing resources and the fact that many infected patients may remain asymptomatic or do not require observed healthcare utilization. There is temporal variability in the extent to which underdiagnosis/underreporting may impact case rates as countries develop different national testing strategies, diagnosis guidance and the sensitivity of utilized testing instruments (e.g. nasopharyngeal versus saliva swabs, antibody testing). The case definition will inherently vary over calendar time and may require, as time continues, adjustment for known issues in case classification. The use of seasonal vaccinations for influenza strains may contribute to disruption in the natural history of influenza in susceptible subgroups. In the absence of vaccination, the adoption and implementation of social distancing measures to isolate and contain SARS-COV-2 spread may also disrupt the natural history of COVID-19. In both cases the success of preventative measures may lead to lower case rates than unmitigated populations though the impact is largely negligible as large populations still remain at-risk for potential infection.

Furthermore, classification of the cases could also vary with respect to calendar time, since underdiagnosis could become more frequent over the course of the pandemic, with positive cases tending to be identified mostly in severe cases. To address this inherent limitation, the study also includes a hospitalisation-based COVID-19 case definition which will likely be more consistent across space and time and better classified for the more serious cases of disease in these data, to provide additional context.

Medical conditions may be underestimated as they will be based on the presence of condition codes, with the absence of such a record taken to indicate the absence of a disease. Meanwhile, medication records indicate that an individual was prescribed or dispensed a particular drug, but this does not necessarily mean that an individual took the drug as originally prescribed or dispensed. Our study could be subject to exposure misclassification with false positives if a patient had a dispensing but did not ingest the drug, but may also be subject to false negatives for non-adherent patients who continued their medication beyond the days supply due to stockpiling. Medication use estimates on the date of hospitalization is particularly sensitive to misclassification, and may conflate baseline concomitant drug history with immediate treatment upon admission. Comparisons of individuals with COVID-19 with individuals previously with influenza has limitations. In particular, observed differences may be explained by changes in clinical practice or data capture procedures over time, rather than by differences in the individuals themselves. This is likely a particular relevant drawback for any comparison of medication use. There may be less time-to-event data in the more emergent population at-risks this will inevitably become a non-issue as the pandemic continues on and more follow-up time is captured. In the immediate analysis, we will investigate the potential variation in complete versus incomplete follow-up time.

# 10. Protection of Human Subjects

The study uses only de-identified data. Confidentiality of patient records will be maintained at all times. Data custodians will remain in full control of executing the analysis and packaging results. There will be no transmission of patient-level data at any time during these analyses. Only aggregate statistics will be captured. Study packages will contain minimum cell count parameters to obscure any cells which fall below allowable reportable limits. All study reports will contain aggregate data only and will not identify individual patients or physicians.

# 11. Management and Reporting of Adverse Events and Adverse Reactions

This study will provide a descriptive summary of individuals at time of diagnosis/test for COVID-19. Adverse events after the first 30 days of a COVID-19 record are outside the scope of the study.

# 12. Plans for Disseminating and Communicating Study Results

The results will be used across multiple papers by the target cohorts, by stratification features, and/or by baseline characteristics, treatments, or outcomes. All results will be posted on the OHDSI website (evidence.ohdsi.org) after completion of the study. At least one paper per topic presented in Section 2.1 describing the study and its results will be written and submitted for publication to a peer-reviewed scientific journal. The results will also be presented at the OHDSI in-person events.

# 

# References

1. Wang D, Hu B, Hu C, *et al.* Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA* 2020;**323**:1061–9. doi:10.1001/jama.2020.1585
2. Arentz M, Yim E, Klaff L, *et al.* Characteristics and Outcomes of 21 Critically Ill Patients With COVID-19 in Washington State. *JAMA* Published Online First: 19 March 2020. doi:10.1001/jama.2020.4326
3. Parag Goyal P, Choi JJ, Pinheiro LC, et al. Clinical Characteristics of Covid-19 in New York City. *NEJM.* 17April 2020. doi: 10.1056/NEJMc2010419
4. Jon Cohen, Kai Kupferschmidt, Countries test tactics in ‘war’ against COVID-19. Science. 2020. Vol. 367, Issue 6484, pp. 1287-1288. 10.1126/science.367.6484.1287
5. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. Journal of the American Medical Informatics Association : JAMIA. 2012;19(1):54-60
6. Voss EA, Makadia R, Matcho A, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inf Assoc 2015;22:553–64. doi:10.1093/jamia/ocu023

# ***Appendix 1: Cohort Definitions (To be updated after phenotype Review is complete)***

**Persons hospitalized with COVID-19 narrow, w/ no prior observation required**

**(**<https://atlas.ohdsi.org/#/cohortdefinition/108>)

Persons hospitalized with COVID-19, age>=18, no prior observation required, narrowly defined = (condition or measurement with positive value) in 21d prior or during (up to date of discharge) of hospitalization, with no COVID associated hospitalization in prior 6 mo, and after 1Dec2019

Initial Event Cohort

People having any of the following:

* a visit occurrence of [OHDSI Covid19 v1] Inpatient Visit5
  + occurrence start is after 2019-12-01

with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: **all events per person.**

For people matching the Primary Events, include:

Having any of the following criteria:

* at least 1 occurrences of a condition occurrence of COVID-19 (including asymptomatic)1

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

* or at least 1 occurrences of a condition occurrence of Any Condition
  + Condition Source Concept is COVID-19 source codes2

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

* or at least 1 occurrences of a measurement of COVID-19 specific testing (pre-coordinated Measurements) - Positive4

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

* or at least 1 occurrences of a measurement of COVID-19 specific testing (pre-coordinated Measurements excluded)3
  + value as concept is any of: Detected, Detected, Positive, Present, Present, Positive

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

* or at least 1 occurrences of an observation of COVID-19 specific testing (pre-coordinated Measurements excluded)3
  + value as concept is any of: Detected, Detected, Positive, Positive, Present, Present

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

* or at least 1 occurrences of an observation of Any Observation

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

Limit cohort of initial events to: **all events per person.**

Inclusion Rules

Inclusion Criteria #1: age >=18

Having all of the following criteria:

* with the following event criteria:
  + with age >= 18

Inclusion Criteria #2: does not have hospitalization for COVID19 in the 6 months preceding admission

Having all of the following criteria:

* exactly 0 occurrences of a visit occurrence of [OHDSI Covid19 v1] Inpatient Visit5

Having any of the following criteria:

* + - at least 1 occurrences of a condition occurrence of COVID-19 (including asymptomatic)1

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

* + - or at least 1 occurrences of a condition occurrence of Any Condition
      * Condition Source Concept is COVID-19 source codes2

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

* + - or at least 1 occurrences of an observation of Any Observation

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

* + - or at least 1 occurrences of a measurement of COVID-19 specific testing (pre-coordinated Measurements) - Positive4

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

* + - or at least 1 occurrences of a measurement of COVID-19 specific testing (pre-coordinated Measurements excluded)3
      * value as concept is any of: Detected, Detected, Positive, Positive, Present, Present

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

* + - or at least 1 occurrences of an observation of COVID-19 specific testing (pre-coordinated Measurements excluded)3
      * value as concept is any of: Detected, Detected, Positive, Positive, Present, Present

where event starts between 21 days Before and all days After index start date and event starts between all days Before and 0 days After index end date

where event starts between 180 days Before and 1 days Before index start date

Limit qualifying cohort to: **all events per person.**

End Date Strategy

Date Offset Exit Criteria

This cohort definition end date will be the index event's end date plus 0 days

Cohort Collapse Strategy:

Collapse cohort by era with a gap size of 0 days.

**Concept Set: COVID-19 (including asymptomatic)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| 37311061 | Disease caused by severe acute respiratory syndrome coronavirus 2 | Condition | SNOMED | NO | YES | NO |

**Concept Set: COVID-19 source codes**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| 586414 | Novel coronavirus infection | Condition | KCD7 | NO | NO | NO |
| 586415 | Provisional assignment of new diseases or emergency use | Condition | KCD7 | NO | NO | NO |
| 710155 | COVID-19 pneumonia (machine translation) | Observation | ICD10CN | NO | NO | NO |
| 710156 | COVID-19 pneumonia (machine translation) | Observation | ICD10CN | NO | NO | NO |
| 710157 | Suspected case of COVID-19 (machine translation) | Condition | ICD10CN | YES | NO | NO |
| 710158 | COVID-19 (machine translation) | Observation | ICD10CN | NO | NO | NO |
| 710159 | Confirmed COVID-19, excluding pneumonia (machine translation) | Observation | ICD10CN | NO | NO | NO |
| 710160 | Emergency use of U07.1 | Disease caused by severe acute respiratory syndrome coronavirus 2 | Observation | ICD10CN | NO | NO | NO |
| 42501115 | Emergency use of U07.1 | Disease caused by severe acute respiratory syndrome coronavirus 2 | Condition | KCD7 | NO | NO | NO |
| 45542411 | Contact with and (suspected) exposure to other viral communicable diseases | Observation | ICD10CM | YES | NO | NO |
| 45600471 | Other coronavirus as the cause of diseases classified elsewhere | Condition | ICD10CM | NO | NO | NO |

**Concept Set: COVID-19 specific testing (pre-coordinated Measurements excluded)**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| 756055 | Measurement of severe acute respiratory syndrome coronavirus 2 | Measurement | OMOP Extension | NO | YES | NO |
| 37310281 | 2019 novel coronavirus not detected | Measurement | SNOMED | YES | YES | NO |
| 37310282 | 2019 novel coronavirus detected | Measurement | SNOMED | YES | YES | NO |

**Concept Set: COVID-19 specific testing (pre-coordinated Measurements) - Positive**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| 37310282 | 2019 novel coronavirus detected | Measurement | SNOMED | NO | YES | NO |

**Concept Set: [OHDSI Covid19 v1] Inpatient Visit**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Concept Id** | **Concept Name** | **Domain** | **Vocabulary** | **Excluded** | **Descendants** | **Mapped** |
| 262 | Emergency Room and Inpatient Visit | Visit | Visit | NO | YES | NO |
| 9201 | Inpatient Visit | Visit | Visit | NO | YES | NO |