Unraveling COVID-19: a large-scale characterization of 4.5 million COVID-19 cases using CHARYBDIS

Daniel Prieto-Alhambra (daniel.prietoalhambra@ndorms.ox.ac.uk)
Centre for Statistics in Medicine (CSM), Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDROMS), University of Oxford, UK https://orcid.org/0000-0002-3950-6346

Kristin Kostka
Real World Solutions, IQVIA, Cambridge, MA, USA https://orcid.org/0000-0003-2595-8736

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 , University of Oxford https://orcid.org/0000-0003-1202-9153

Anthony Sena
Janssen R&D, Titusville NJ, USA, 2) Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands https://orcid.org/0000-0001-8630-5347

Andrea Pistillo
Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain

Sara Khalid
Centre for Statistics in Medicine, NDORMS, University of Oxford, UK

Lana Lai
Division of Cancer Sciences, School of Medical Sciences, University of Manchester, UK

Asieh Golozar
Regeneron Pharmaceuticals, NY USA, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, MD USA https://orcid.org/0000-0002-4243-155X

Thamir M Alshammari
Medication Safety Research Chair, King Saud University, Riyadh, Saudi Arabia

Dalia Dawoud
National Institute for Health and Care Excellence, London, UK

Fredrik Nyberg
School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden https://orcid.org/0000-0003-0892-5668

Adam Wilcox
Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA, 2) UW Medicine, Seattle, WA, USA

Alan Andryc
Janssen R&D, Titusville NJ, USA
Andrew Williams
Tufts Institute for Clinical Research and Health Policy Studies, US https://orcid.org/0000-0002-0692-412X
Anna Ostropolets
Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032, USA
Carlos Areia
Nuffield Department of Clinical Neurosciences, University of Oxford, UK
Chi Young Jung
Division of Respiratory and Critical Care Medicine, Department of Internal Medicine, Daegu Catholic University Medical Center, Daegu, Korea
Christopher Harle
University of Florida Health, Gainesville, FL, USA
Christian Reich
Real World Solutions, IQVIA, Cambridge, MA, USA
Clair Blacketer
Janssen R&D, Titusville NJ, USA, 2) Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
Daniel Morales
Division of Population Health and Genomics, University of Dundee, UK
David A. Dorr
Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA
Edward Burn
Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain, https://orcid.org/0000-0002-9286-1128
Elena Roel
https://orcid.org/0000-0002-1964-3546
Eng Hooi Tan
Centre for Statistics in Medicine, NDORMS, University of Oxford, UK
Evan Minty
O'Brien Institute for Public Health, Faculty of Medicine, University of Calgary, Canada
Frank DeFalco
Janssen R&D, Titusville NJ, USA
Gabriel de Maeztu
IOMED, Barcelona, Spain
Gigi Lipori

University of Florida Health

Heba Alghoul

Faculty of Medicine, Islamic University of Gaza, Palestine https://orcid.org/0000-0001-8234-5843

Hong Zhu

Nanfang Hospital, Southern Medical University, Guangzhou, China

Jason Thomas

Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA https://orcid.org/0000-0003-3892-7197

Jiang Bian

University of Florida https://orcid.org/0000-0002-2238-5429

Jimyung Park

Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea

Jordi Martínez Roldán

Director of Innovation and Digital Transformation, Hospital del Mar, Barcelona, Spain

Jose Posada

Stanford University School of Medicine, Stanford, California, USA https://orcid.org/0000-0003-3864-0241

Juan M Banda

Georgia State University, Department of Computer Science, Atlanta, GA, USA

Juan P Horcajada

Department of Infectious Diseases, Hospital del Mar, Institut Hospital del Mar d'Investigació Mèdica (IMIM), Universitat Autònoma de Barcelona. Universitat Pompeu Fabra, Barcelo

Julianna Kohler

United States Agency for International Development, Washington, DC, USA

Karishma Shah

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK

Karthik Natarajan

Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032, USA, 2) New York-Presbyterian Hospital, 622 W 168 St, PH20 New York, NY 10032 USA https://orcid.org/0000-0002-9066-9431

Kristine Lynch

VINCI, VA Salt Lake City Health Care System, Salt Lake City, VA, & Division of Epidemiology, University of Utah, Salt Lake City, UT

Li Liu

Biomedical Big Data Center, Nanfang Hospital, Southern Medical University, Guangzhou, China

Lisa Schilling

Data Science to Patient Value Program, University of Colorado Anschutz Medical Campus

Martina Recalde

Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain

Matthew Spotnitz

Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032, USA

Mengchun Gong

DHC Technologies Co. Ltd, Beijing, China

Michael Matheny

VINCI, Tennessee Valley Healthcare System VA, Nashville, TN & Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN

Neus Valveny

Real-World Evidence, TFS, Barcelona, Spain

Nicole Weiskopf

Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA

Nigam Shah

Stanford University https://orcid.org/0000-0001-9385-7158

Osaid Alser

Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Paula Casajust

Trial Form Support https://orcid.org/0000-0003-2733-5436

Rae Woong Park

Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Korea

Robert Schuff

Knight Cancer Institute, Oregon Health & Science University

Sarah Seager

Real World Solutions, IQVIA, Cambridge, MA, USA

Scott DuVall

VA Informatics and Computing Infrastructure, VA Salt Lake City Health Care System, Salt Lake City, UT, USA

Seng Chan You

Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea

Seokyoung Song

Department of Anesthesiology and Pain Medicine, Catholic University of Daegu, School of Medicine, Daegu, Korea

Sergio Fernández-Bertolín

Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain

Stephen Fortin

Observational Health Data Analytics, Janssen Research and Development, Raritan, NJ, USA

Tanja Magoc

University of Florida Health

Thomas Falconer

Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032, USA

Vignesh Subbian

College of Engineering, The University of Arizona, Tucson, Arizona, USA

Vojtech Huser

National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

Waheed-Ul-Rahman Ahmed

Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK, 2) College of Medicine and Health, University of Exeter, St Luke's Campus, E

https://orcid.org/0000-0003-0880-0355

William Carter

Data Science to Patient Value Program, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

Yin Guan

DHC Technologies Co. Ltd, Beijing, China

Yankuic Galvan

University of Florida Health

Xing He

University of Florida Health

Peter Rijnbeek

Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

George Hripcsak

Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032, USA, 2) New York-Presbyterian Hospital, 622 W 168 St, PH20 New York, NY 10032 USA

Patrick Ryan

Janssen R&D

Marc Suchard

Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles

https://orcid.org/0000-0001-9818-479X

Biological Sciences - Article

Keywords: COVID-19, OHDSI, OMOP CDM, hospital admission, descriptive epidemiology, real world data, real world evidence, open science

DOI: https://doi.org/10.21203/rs.3.rs-279400/v1

Abstract

Background: Routinely collected real world data (RWD) have great utility in aiding the novel coronavirus disease (COVID-19) pandemic response [1,2]. Here we present the international Observational Health Data Sciences and Informatics (OHDSI) [3] Characterizing Health Associated Risks, and Your Baseline Disease In SARS-COV-2 (CHARYBDIS) framework for standardisation and analysis of COVID-19 RWD.

Methods: We conducted a descriptive cohort study using a federated network of data partners in the United States, Europe (the Netherlands, Spain, the UK, Germany, France and Italy) and Asia (South Korea and China). The study protocol and analytical package were released on 11th June 2020 and are iteratively updated via GitHub [4].

Findings: We identified three non-mutually exclusive cohorts of 4,537,153 individuals with a clinical *COVID-19 diagnosis or positive test,* 886,193 *hospitalized with COVID-19*, and 113,627 *hospitalized with COVID-19 requiring intensive services*. All comorbidities, symptoms, medications, and outcomes are described by cohort in aggregate counts, and are available in an interactive website: https://data.ohdsi.org/Covid19CharacterizationCharybdis/.

Interpretation: CHARYBDIS findings provide benchmarks that contribute to our understanding of COVID-19 progression, management and evolution over time. This can enable timely assessment of real-world outcomes of preventative and therapeutic options as they are introduced in clinical practice.

Introduction

The World Health Organization (WHO) declared the coronavirus disease 2019 (COVID-19) pandemic on 11 March 2020 after 118,000 reported cases in over 110 countries [5]. By 2021, the number of COVID-19 cases has increased to over 90,000,000 globally, and as we write the death toll has reached 2 million [6]. Thousands of publications have attempted to aid our scientific understanding of this public health emergency [7,8].

Routinely collected real world data (RWD) are a powerful asset for an evolving pandemic response [1,2]. Each data source provides novel information, be it the geographic variability of COVID-19, the impact of varying government strategies to contain spread or the evolution of treatment protocols. With extensive heterogeneity in public health strategies and clinical care across the world [9], a large repeated multi-center study to describe disease across locations, practices, and populations, but that holds data analysis constant would go far in determining what factors impact observed differences.

RWD networks are vital in helping to understand the magnitude of the problem, and developing possibly mitigating strategies both globally and locally [10,11]. Here we present the global Observational Health Data Sciences and Informatics (OHDSI) community response to the COVID-19 pandemic [3]. Founded in 2015, the OHDSI data network enabled a rapid baseline understanding of COVID-19 in emerging hotspots (United States of America [USA], Spain and South Korea) [12]. Our work evolved into a systematic

framework for analysing and reporting COVID-19 RWD that we call Characterizing Health Associated Risks, and Your Baseline Disease In SARS-COV-2 (CHARYBDIS).

CHARYBDIS offers multiple insights into COVID-19 clinical presentations, management and progression. We set out to continually describe baseline demographics, clinical characteristics, treatments received, and outcomes among individuals diagnosed and hospitalized with COVID-19 in actual practice settings in nine countries from three continents. Our body of research is a freely available, foundational result set that can provide benchmarks in how COVID-19 manifests over time including its inevitable evolution as we roll-out vaccines and treatments.

Results

All comorbidities, presenting symptoms, medications and outcomes are reported by each cohort in aggregate counts, and are available in an interactive website: https://data.ohdsi.org/Covid19CharacterizationCharybdis/.

Patient characteristics

Overall, we identified three non-mutually exclusive cohorts of 4,537,153 individuals with a clinical *COVID-19 diagnosis or positive test,* 886,193 *hospitalized with COVID-19*, and 113,627 *hospitalized with COVID-19 requiring intensive services* (Figure 1). Of these, the cohorts including patients with the requirement of at least of 365 days before index: 3,279,518 with a clinical *COVID-19 diagnosis or laboratory positive test,* 636,810 *hospitalized with COVID-19*, and 63,636 *hospitalized with COVID-19 requiring intensive services* (Supplementary Tables 3 & 4).

Geographic distribution

The USA data partners contributed 96% of the *diagnosed with COVID-19 cohorts*, including the single largest diagnosed cohort from IQVIA Open Claims (n=2,785,812). Europe contributed 4% of the *diagnosed with COVID-19 cohorts,* owing the single largest regional diagnosed cohort to SIDIAP-Spain (n=124,305). Asia contributed less than 1% of *diagnosed with COVID-19 cohorts*, with the single largest regional diagnosed cohort contributed from Daegu Catholic University Medical Center (n=599).

Demographic distribution

In the USA, the proportion of diagnosed cases generally decreased with age, with most diagnosed cases being within the 25 to 60 age group. The proportion of cases hospitalized and intensive services increased with age, with the highest proportions of cases of hospitalized, or intensive cases in the 60 to 80 year age group (Figure 2). A slightly higher proportion of women were diagnosed than men but a greater proportion of men were hospitalized (and where available, required intensive services) than women in the USA databases. In Europe, databases captured diagnosed or hospitalised cohorts but had limited information on intensive services. In Europe, databases capturing hospitalized cases (HMAR, HM-Hospitales, SIDIAP, and SIDIAP-H) showed a similar trend to the USA databases in that there was a higher

proportion of men were hospitalized than women (Supplementary Figure 1). Unlike the USA and European databases, there was also a higher proportion of women in hospitalized cases in the South Korean database (HIRA). Age-wise trends in the European and Asian databases were similar to those in the USA databases, in that the bulk of the diagnosed cases were in the 25 to 60 year age group, whilst the majority of the hospitalized cases were in the 60 to 80 year age group (Supplementary Figure 1).

Comorbidities

Overall, the proportion of patients with type 2 diabetes mellitus, hypertension, chronic kidney disease, end stage renal disease, heart disease, malignant neoplasm, obesity, dementia, auto-immune condition, chronic obstructive pulmonary disease (COPD), and asthma was higher in the hospitalised cohort as compared to the diagnosed (Tables 1 and 2). Data on tuberculosis, human immunodeficiency viruses (HIV), and hepatitis C infections were sparse, and where available the proportions were generally low (<=1%). In the US databases, the proportion of pregnant women was generally higher in the hospitalised cohort than in the diagnosed, but not so in two European databases (HM and SIDIAP). The remaining five European and one of the Asian databases had data on pregnant women only in the hospitalised cohort, the proportion of which was < 2%.

Other analyses

Dyspnea, cough, and fever were the most common symptoms in diagnosed and hospitalized cohorts (Supplementary Table 5). Where recorded, the proportion of dyspnea and malaise/fatigue was consistently higher in the hospitalised cohort as compared to the diagnosed. Anosmia/hyposmia/dysgeusia was present in less than 1% individuals in all but one database and more common in the diagnosed than the hospitalised cohorts.

We further described a total of 19,222 conditions and 2,973 medications registered during the year prior to the index date (Supplementary Figure 2). The same information is also described for 30 days prior to the index date, at index date, or during the first 30 days after index date (this can be explored in detail at https://data.ohdsi.org/Covid19CharacterizationCharybdis/).

Discussion

Summary of key findings

We described characteristics of 4,537,153 individuals with a clinical *COVID-19 diagnosis or positive test,* 886,193 *hospitalized with COVID-19*, and 113,627 *hospitalized with COVID-19 requiring intensive services* from 9 countries. Up to 22,200 unique aggregate characteristics have been produced across databases, with all made publicly available in an accompanying website. The cumulative evidence obtained from different regions and at different points in the pandemic can guide in 1) better patient characterization and stratification, 2) identifying areas of gap in knowledge/evidence, and 3) generating hypotheses for future research.

Findings in context

In April 2020, the National COVID Cohort Collaborative (N3C) chose the OMOP CDM as the data model for centralizing patient-level data to study patterns in COVID-19 patients [20]. This network has over 80 participating institutions and is enabling many US institutions in adoption of common data models in COVID-19 research. This program has two major differences: 1) data are limited to US only sites and 2) the centralized data approach requires significant programmatic oversight. In contrast to this and other notable RWD initiatives, CHARYBDIS uses an existing decentralized network, open to all, with no requirement to move patient-level data [21]. This enables the opportunity to integrate results from regions within more restrictive data sharing policies, such as Europe and Asia.

The Consortium for Clinical Characterization of COVID-19 by EHR (4CE), is another multi-site data-sharing collaborative of 342 hospitals in the US and in Europe, utilizing i2b2 or OMOP data models [22]. Despite its extensive footprint, 4CE cohorts remain smaller than the scope of CHARYBDIS with only 36,447 patients with COVID-19 as of August 2020 [22]. Even with cohort overlap, our work to date with CHARYBDIS is substantial spanning 4.5 million COVID-19 patients across three continents.

The “tragic data gap” undermining response to the pandemic [23] is effected by inadequate utilization of and access to high-quality RWD. Large scale initiatives like CHARYBDIS can offer critical infrastructure for mobilizing simple descriptive epidemiological studies that are fundamentally important in tracking the evolution and ultimate management of this pandemic. Our findings can help proivde context on where to direct future funding and carry out additional research. The information generated from CHARYBDIS can inform the response to the pandemic, including both public health restrictions (non-pharmacological interventions) and vaccination strategies worldwide. As we continue our research, we are also actively curating relationships with data partners to drive inpatient-outpatient linkages and understand COVID-19 patient trajectories across care settings.

Study strengths

Our study has several strengths. This study is unique in its approach to characterizing COVID-19 cases across an international network of healthcare systems with varied policies enacted to combat this pandemic. This allows better understanding of the implications of the pandemic for different countries and regions, in the context of an international comparison. Particularly, it provides visibility into the inherent variability of patient characteristics across healthcare settings. This study is the most comprehensive federated network of healthcare sites in the world, creating the single largest cohort study on diagnosed and hospitalized COVID-19 cases to date. The large, diverse sample size allows also for the identification of populations of great interest, including children and adolescents, pregnant women, patients with a history of cancer, or patients with HIV, who were also infected with COVID-19, and who will be the focus of in-depth future investigations.

Study limitations

We recognize there are limitations in our approach. First, this study is descriptive in nature and was not designed for causal inference. The observed differences between groups (e.g. diagnosed versus hospitalized) should therefore not be interpreted as causal effects. Answering causal questions is especially difficult in COVID-19 because of the varying processes by which patients were screened, tested, admitted, and treated; the critical importance of knowing the exact timing of treatments and outcomes in severe cases; and the lack of appropriate comparison groups. Simple multivariable models by themselves will not sufficiently address bias for multiple questions and were purposely not applied here. This study was carried out using data recorded in routine clinical practice and based on electronic health records (EHRs) and/or claims data. The analysed data are therefore expected to be incomplete in some respects and may have erroneous entries, leading to potential misclassification. We have selectively reported database-specific outcomes to minimise the impact of incompleteness. Additionally, the under-reporting of symptoms observed in these data is a key finding of this study, and should be taken into consideration in previous and future similar reports from 'real world' cohorts. Differential reporting in different databases is likely a function of differential coding practice as well as of variability in disease severity, with milder/less symptomatic cases more likely presenting in outpatient and primary care EHR, and more severe ones in hospital databases. Finally, the current result submissions are prejudiced to data in the initial wave of COVID-19 cases and may not be representative of the data during subsequent waves. We currently lack data partners in low to middle income countries and are actively building collaborations in these areas. As data are accumulated over time, future updates of the results will provide the opportunity to study more recent cohorts of COVID-19 patients, who seem to have a better prognosis overall compared to those diagnosed in the first half of the year.

Conclusions

We present the foundation for an epidemiological framework to perform large scale characterization of the presentation, management, and outcomes of COVID-19 as observed in actual practice settings worldwide. We have characterized the natural history of over 4.5 million COVID-19 patients from the USA, 6 European countries and 2 Asian countries. This work allows deep phenotyping of COVID-19, serving as a repeatable, reproducible method to capture the evolving natural history of this novel coronavirus and can be extended to future pandemics. Leveraging our global federated network to corroborate single center findings can provide context to national database findings in the presence of regional variability in COVID-19 policies. This effort provides critical infrastructure for mobilizing descriptive studies that are fundamentally important in tracking the evolution and ultimate management of this pandemic.

Methods

Study design, setting and data sources

We conducted a descriptive cohort study using a federated network of data partners in the USA, Europe (the Netherlands, Spain, the UK, Germany, France and Italy) and Asia (South Korea and China). We required each data partner to map their source system to the Observational Medical Outcomes

Partnership (OMOP) common data model (CDM) [13–15]. The use of a CDM ensured shared conventions, including consistent representation of clinical terms across coding systems. We deployed a common data quality tool for repeated assessment and monitoring the adherence to conventions across the network [16,17]. We ensured reproducibility by using the same package of analytical code for all contributing data partners [18].

The study protocol and analytical package were released on 11 June 2020 and iterative updates have continued to be released via GitHub: https://github.com/ohdsi-studies/Covid19CharacterizationCharybdis [4]. As of February 2021, 26 databases have contributed to the CHARYBDIS study (Supplementary Table 1). Contributing institutes ranged from major academic medical centers to small community hospitals from across three continents. While most data were captured from March to June 2020, a subset of data partners submitted updates through October 2020. Two sites report data through December 2020. Additional updates are expected as data partners refresh their OMOP CDM data. Prior to performing these analyses, all the data partners obtained Institutional Review Board (IRB) or equivalent governance approval. Each data partner executed the study package locally on their OMOP CDM. Only aggregate results from each database were publicly shared. Minimum cell sizes were determined by institutional protocols. All data partners consented to the external sharing of the result set on data.ohdsi.org.

Study population and follow-up

We focused on three non-mutually exclusive COVID-19 cohorts: i) *diagnosed with COVID-19* (a positive SARS-CoV-2 laboratory test or clinical diagnosis of COVID-19 - earliest event served as the index date); ii) *hospitalized with COVID-19* and; iii) *hospitalized with COVID-19 and requiring intensive services*. The codes used to identify cohorts and more detail on the definitions of the above cohorts can be found in Supplementary Table 2. These cohorts were generated both with a requirement of at least 365 days of data availability prior to the index date, and without any requirement for prior observation time. Datamarts created specifically for COVID-19 tracking may be unable to support extensive lookback periods and thus, we used multiple definitions to ensure inclusiveness in our approach. Cohorts were followed from their cohort-specific index date to the earliest of death, end of the observation period, and up to 30 days post-index.

Stratifications

Each cohort was analyzed by the overall study population and stratified by additional available characteristics including: follow-up time; socio-demographics, baseline comorbidities, pregnancy status (yes/no), and flu-like symptom episodes (yes/no). Detailed definitions of each stratification are available in Supplementary Table 2.

Baseline characteristics, symptoms, medication use and outcomes of interest

Information on socio-demographics was identified at or before baseline (index date). All conditions, symptoms and medications were identified and described at four different time intervals (1 year prior, 30 days prior, at index and up to 30 days after index). The definition of each symptom and outcome is provided in Supplementary Table 2.

Statistical analysis

We built this analysis using Health Analytics Data-to-Evidence Suite (HADES), a set of open source R packages for large scale analytics [19]. Proportions, standard deviations (SD), and standardized mean differences (SMD) within each subgroup were tabulated as pre-specified in our study protocol. This analysis was descriptive in nature with no causal inference intended. Only cohorts or stratified sub-cohorts with a minimum sample size of 140 subjects were characterized. This cut-off was deemed necessary to estimate with sufficient precision the prevalence of a previous condition or 30-day risk of an outcome affecting >=10% of the study population. SMDs were plotted in Manhattan-style plots, a type of scatter plot designed to visualize large data with a distribution of higher-magnitude values. Scatter plots were also created to compare the described conditions, symptoms and demographics of patients diagnosed (Y axis) to those hospitalized (X axis) with COVID-19.

Declarations

Funding

The European Health Data & Evidence Network has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. This research received partial support from the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), US National Institutes of Health, US Department of Veterans Affairs, the Health Department from the Generalitat de Catalunya with a grant for research projects on SARS-CoV-2 and COVID-19 disease organized by the Direcció General de Recerca i Innovació en Salut, Janssen Research & Development, and IQVIA. The University of Oxford received funding related to this work from the Bill & Melinda Gates Foundation (Investment ID INV-016201 and INV-019257). This study was supported by National Key Research & Development Program of China (Project No.2018YFC0116901). OHSU received support from Gates Foundation, INV-016910 and the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR002369. The University of Washington received a grant related to this work from the Bill & Melinda Gates Foundation (INV-016910). No funders had a direct role in this study. The views and opinions expressed are those of the authors and do not necessarily reflect those of the Clinician Scientist Award programme, NIHR, Department of Veterans Affairs or the United States Government, NHS, National Institute for Health and Care Excellence (NICE) or the Department of Health, England.

Ethical approval

All the data partners received Institutional Review Board (IRB) approval or exemption. STARR-OMOP had approval from IRB Panel #8 (RB-53248) registered to Leland Stanford Junior University under the Stanford Human Research Protection Program (HRPP). The use of VA data was reviewed by the Department of Veterans Affairs Central IRB, was determined to meet the criteria for exemption under Exemption Category 4(3), and approved for Waiver of HIPAA Authorization. The research was approved by the Columbia University Institutional Review Board as an OHDSI network study. The use of SIDIAP was approved by the Clinical Research Ethics Committee of the IDIAPJGol (project code: 20/070-PCV). The use of HMAR was approved by the Parc de Salut Mar Clinical Research Ethics Committee**.** The use of CPRD was approved by the Independent Scientific Advisory Committee (ISAC) (protocol number 20_059RA2). This study is approved by the University of Florida IRB under protocol IRB202100175. Some databases used (HealthVerity, Premier, IQVIA Open Claims, Optum EHR, and Optum SES) in these analyses are commercially available, syndicated data assets that are licensed by contributing authors for observational research. These assets are de-identified commercially available data products that could be purchased and licensed by any researcher. The collection and de-identification of these data assets is a process that is commercial intellectual property and not privileged to the data licensees and the co-authors on this study. Licensees of these data have signed Data Use Agreements with the data vendors which detail the usage protocols for running retrospective research on these databases. All analyses performed in this study were in accordance with Data Use Agreement terms as specified by the data owners. As these data are deemed commercial assets, there is no Institutional Review Board applicable to the usage and dissemination of these result sets or required registration of the protocol with additional ethics oversight. Compliance with Data Use Agreement terms, which stipulate how these data can be used and for what purpose, is sufficient for the licensing commercial entities. Further inquiry related to the governance oversight of these assets can be made with the respective commercial entities: HealthVerity (healthverity.com), Premier (premierinc.com), IQVIA (iqvia.com) and Optum (optum.com). At no point in the course of this study were the authors of this study exposed to identified patient-level data. All result sets represent aggregate, de-identified data that are represented at a minimum cell size of >5 to reduce potential for re-identification. Furthermore, the New England Institutional Review Board of Janssen Research & Development (Raritan, NJ) has determined that studies conducted on licensed copies of Premier, Optum EHR, Optum SES and HealthVerity are exempt from study-specific IRB review, as these studies do not qualify as human subjects research.

Competing interest statement

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare:

Ms. Kostka is an employee of IQVIA. Mr. Sena is an employee and holds stock at Janssen Research & Development, a Johnson and Johnson family of companies. Dr. Golozar reports personal fees from Regeneron Pharmaceuticals, outside the submitted work. She is a full-time employee at Regeneron Pharmaceuticals. This work was not conducted at Regeneron Pharmaceuticals. Dr. Nyberg reports other funding from AstraZeneca, outside the submitted work. Dr. Wilcox reports grants from Bill and Melinda

Gates Foundation, grants from National Institute of Health, during the conduct of the study Mr. Andryc is an employee of Janssen Research & Development, a subsidiary of Johnson & Johnson. Dr. Reich is an employee of IQVIA. Dr. Blacketer reports she is an employee and holds stock at Janssen Research & Development, a Johnson and Johnson family of companies. Dr. Morales is supported by a Wellcome Trust Clinical Research Development Fellowship (Grant 214588/Z/18/Z) and reports grants from Chief Scientist Office (CSO), grants from Health Data Research UK (HDR-UK), grants from National Institute of Health Research (NIHR), outside the submitted work. Mr. DeFalco reports he is an employee and holds stock at Janssen Research & Development, a Johnson and Johnson family of companies. Jason Thomas reports grants from Bill and Melinda Gates Foundation, grants from National Institute of Health, during the conduct of the study. Dr. Posada reports grants from National Library of Medicine, during the conduct of the study. Dr. Natarajan reports grants from US NIH, during the conduct of the study. Dr. Matheny reports grants from US NIH, grants from US VA HSR&D, during the conduct of the study. Dr. Weiskopf reports personal fees from Merck, outside the submitted work. Dr. Shah reports grants from National Library of Medicine, during the conduct of the study. Dr. Park reports grants from Ministry of Trade, Industry & Energy, Republic of Korea, grants from Ministry of Health & Welfare, Republic of Korea, grants from Bill & Melinda Gates Foundation, during the conduct of the study. Ms. Seager is an employee of IQVIA. Dr. DuVall reports grants from Anolinx, LLC, grants from Astellas Pharma, Inc, grants from AstraZeneca Pharmaceuticals LP, grants from Boehringer Ingelheim International GmbH, grants from Celgene Corporation, grants from Eli Lilly and Company, grants from Genentech Inc., grants from Genomic Health, Inc., grants from Gilead Sciences Inc., grants from GlaxoSmithKline PLC, grants from Innocrin Pharmaceuticals Inc., grants from Janssen Pharmaceuticals, Inc., grants from Kantar Health, grants from Myriad Genetic Laboratories, Inc., grants from Novartis International AG, grants from Parexel International Corporation through the University of Utah or Western Institute for Veteran Research outside the submitted work. Dr. Fortin is an employee of Janssen R&D, a subsidiary of Johnson and Johnson. Dr. Vignesh reports grants from State of Arizona; Arizona Board of Regents, during the conduct of the study; grants from National Science Foundation, grants from Agency for Healthcare Research and Quality, grants from National Institutes of Health, outside the submitted work; .Dr. Subbian reports grants from State of Arizona; Arizona Board of Regents, during the conduct of the study; grants from National Science Foundation, grants from Agency for Healthcare Research and Quality, grants from National Institutes of Health, outside the submitted work. Dr. Rijnbeek reports grants from Innovative Medicines Initiative, from Janssen Research and Development, during the conduct of the study. Dr. Hripcsak reports grants from US NIH, during the conduct of the study. Dr. Ryan reports and is employee of Janssen Research and Development and shareholder of Johnson & Johnson. Dr. Suchard reports grants from US National Institutes of Health, grants from Department of Veterans Affairs, during the conduct of the study; grants from IQVIA, personal fees from Janssen Research and Development, grants from US Food and Drug Administration, personal fees from Private Health Management, outside the submitted work. Dr. Prieto-Alhambra reports grants and other from AMGEN, grants, non-financial support and other from UCB Biopharma, grants from Les Laboratoires Servier, outside the submitted work; and Janssen, on behalf of IMI-funded EHDEN and EMIF consortiums, and Synapse Management Partners have supported training programmes organised by DPA's department and open for external participants.

The views expressed are those of the authors and do not necessarily represent the views or policy of the Department of Veterans Affairs or the United States Government. No other relationships or activities that could appear to have influenced the submitted work.

Transparency declaration

Lead authors affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

Contributorship statement

KK, TDS, APU, AGS, AP, LL, PC, EB, VH, FN, SK, JK, AG, MAS, PR, GH, MS, AO, SD, MM, LMS, OA, CA, HA, KaS, WurA, JMB, NV, GdM, TMA, PJR, DPA contributed to the conceptualization and design of the study. KK, TDS, APU, AGS, AP, LL, PC, EB, VH, FN, SK, AG, MAS, PR, GH, MS, AO, SD, MM, LMS, NV, GdM, PJR, DPA contributed to the analysis phase of the study. KK, TDS, APU, AGS, AP, PC, SFB, EB, JAT, ABW, SK, PR, GH, TF, KN, AA, SF, NS, JoP, AW, KL, WC, CB, FD, CR, SGY, JyP, RWP, SS, CYJ, HZ, LiL, MG, YG, YZ, PJR, DPA, DavidD, RS, NW, XH, TM, CH, GL, JB, YanG are data owners and contribute to the extract-transform-load of their data to the OMOP CDM and the analytical execution of the study package within their local environments. KK, TDS, APU, AGS, AP, LL, PC, EB, SK, MR, ER, AG, JK, MAS, PR, GH, DD, VS, TMA, EHT, EM, MAS, PJR, DPA were critical to drafting the manuscript and the overall interpreting results.

Acknowledgements

We would like to acknowledge the patients who suffered from or died of this devastating disease, and their families and caregivers. We would also like to thank the social workers and healthcare professionals involved in the management of COVID-19 during these challenging times, from primary care to intensive care units.

Data sharing statement

Analyses were performed locally in compliance with all applicable data privacy laws. Although the underlying identified patient data is not readily available to be shared, authors contributing to this paper have direct access to the data sources used in this study. All results (e.g. aggregate statistics, not presented at a patient-level with redactions for minimum cell count) are available for public inquiry. These results are inclusive of site-identifiers by contributing data sources to enable interrogation of each contributing site. All analytic code and result sets are made available at: https://github.com/ohdsi-studies/Covid19CharacterizationCharybdis

References

1. Kent S, Burn E, Dawoud D, Jonsson P, Østby JT, Hughes N, et al. Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. Pharmacoeconomics.

2020. doi:10.1007/s40273-020-00981-9

2. Forrest CB, McTigue KM, Hernandez AF, Cohen LW, Cruz H, Haynes K, et al. PCORnet® 2020: current state, accomplishments, and future directions. J Clin Epidemiol. 2021;129: 60–67. doi:10.1016/j.jclinepi.2020.09.036
3. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216: 574–578. Available: https://www.ncbi.nlm.nih.gov/pubmed/26262116
4. Sena A, Kostka K, Schuemie M, jdposada. ohdsi-studies/Covid19CharacterizationCharybdis: Charybdis v1.1.1 - Publication Package. 2020. doi:10.5281/zenodo.4033034
5. WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020. [cited 23 Jan 2021]. Available: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020
6. COVID-19 Map - Johns Hopkins Coronavirus Resource Center. [cited 11 Jan 2021]. Available: https://coronavirus.jhu.edu/map.html
7. COVID-19-related medical research: a meta-research and critical appraisal. 5 Jan 2021 [cited 10 Jan 2021]. Available: https://www.docwirenews.com/abstracts/covid-19-related-medical-research-a-meta-research-and-critical-appraisal/
8. Teixeira da Silva JA, Tsigaris P, Erfanmanesh M. Publishing volumes in major databases related to Covid-19. Scientometrics. 2020; 1–12. doi:10.1007/s11192-020-03675-3
9. Subbian V, Solomonides A, Clarkson M, Rahimzadeh VN, Petersen C, Schreiber R, et al. Ethics and Informatics in the Age of COVID-19: Challenges and Recommendations for Public Health Organization and Public Policy. J Am Med Inform Assoc. 2020. doi:10.1093/jamia/ocaa188
10. Madhavan S, Bastarache L, Brown JS, Butte AJ, Dorr DA, Embi PJ, et al. Use of electronic health records to support a public health response to the COVID-19 pandemic in the United States: a perspective from 15 academic medical centers. J Am Med Inform Assoc. 2020. doi:10.1093/jamia/ocaa287
11. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584: 430–436. doi:10.1038/s41586-020-2521-4
12. Burn E, You SC, Sena AG, Kostka K, Abedtash H, Abrahão MTF, et al. Deep phenotyping of 34,128 adult patients hospitalised with COVID-19 in an international network study. Nat Commun. 2020;11: 5009. doi:10.1038/s41467-020-18849-z
13. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19: 54–60. doi:10.1136/amiajnl-2011-000376
14. Ryan PB, Madigan D, Stang PE, Overhage JM, Racoosin JA, Hartzema AG. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. Stat Med. 2012;31: 4401–4415. doi:10.1002/sim.5620

15. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, et al. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc. 2010;17: 652–662. doi:10.1136/jamia.2009.002477
16. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4: 1244. doi:10.13063/2327-9214.1244
17. Observational Health Data Sciences, Informatics. Chapter 15 Data Quality. 11 Jan 2021 [cited 23 Jan 2021]. Available: https://ohdsi.github.io/TheBookOfOhdsi/DataQuality.html#data-quality-in-general
18. Schuemie MJ, Cepeda MS, Suchard MA, Yang J, Tian Y, Schuler A, et al. How Confident Are We about Observational Findings in Healthcare: A Benchmark Study. Harv Data Sci Rev. 2020;2. doi:10.1162/99608f92.147cc28e
19. Observational Health Data Sciences and Informatics. HADES. [cited 11 Jan 2021]. Available: https://ohdsi.github.io/Hades/index.html
20. Haendel M, Chute C, Gersing K. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. J Am Med Inform Assoc. 2020. doi:10.1093/jamia/ocaa196
21. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020;584: 430–436. doi:10.1038/s41586-020-2521-4
22. 4CE Collaborative, Weber GM, Hong C, Palmer NP, Avillach P, Murphy SN, et al. International comparisons of harmonized laboratory value trajectories to predict severe COVID-19: Leveraging the 4CE collaborative across 342 hospitals and 6 countries: A retrospective cohort study. bioRxiv. medRxiv; 2020. doi:10.1101/2020.12.16.20247684
23. Schneider EC. Failing the Test - The Tragic Data Gap Undermining the U.S. Pandemic Response. N Engl J Med. 2020;383: 299–302. doi:10.1056/NEJMp2014836

Tables

Tables 1-2 are available in the Supplementary Files.

Figures

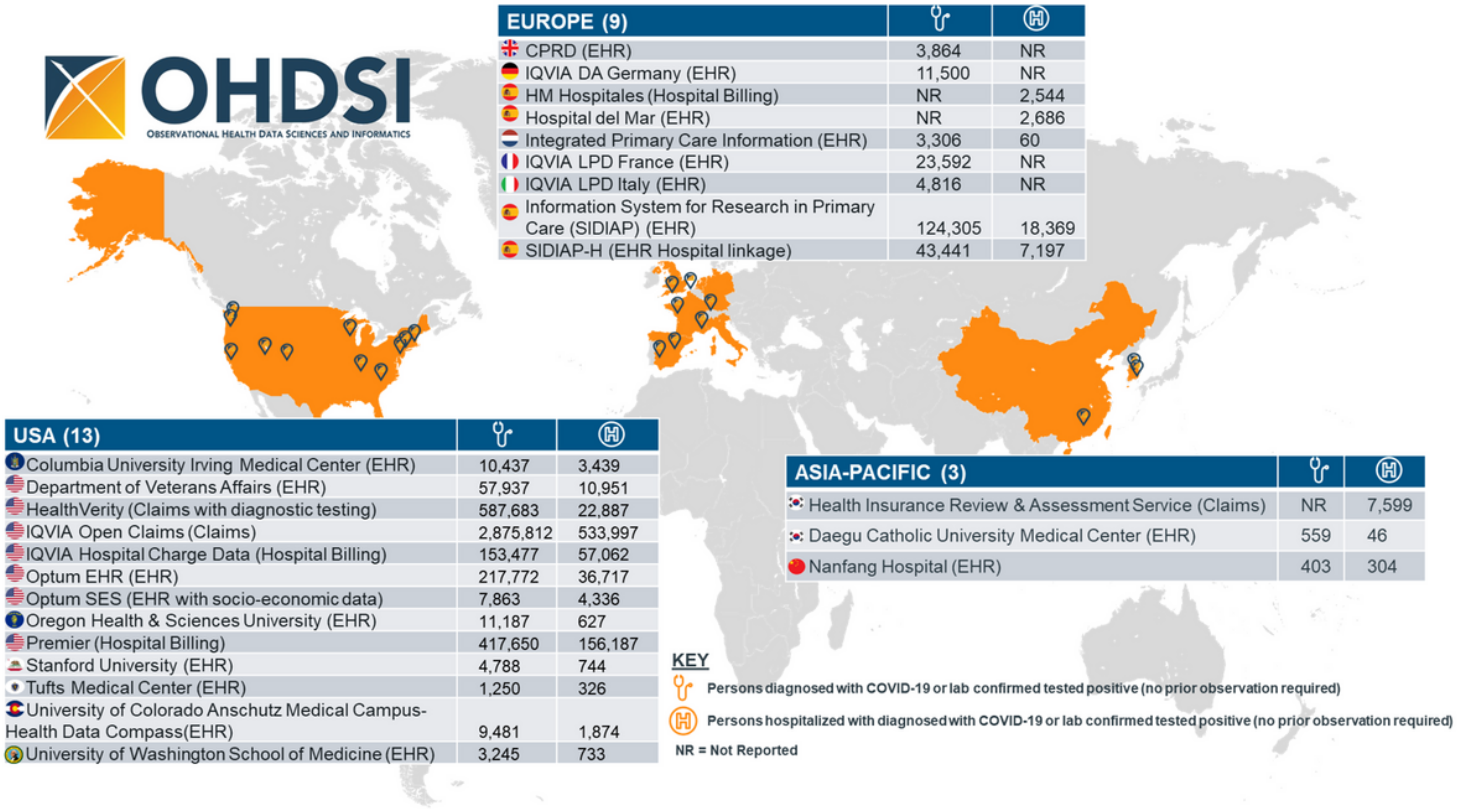

Figure 1

COVID-19 cases across the OHDSI COVID-19 network. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Figure 2

Distribution of diagnosed, hospitalized and requiring intensive services COVID-19 cases by age and sex across the OHDSI COVID-19 network in the United States NB: In each subplot, the x-axis represents what proportion of all women (left) and all men (right) fall in each age category. No prior observation period required in the cohorts shown in this figure. Cohorts must be >=140 people to be reported in this analysis. Abbreviations: diag: diagnosed; hosp: hospitalized; i.s.: hospitalized and requiring intensive services. Abbreviations: CU-AMC-HDC: U of Colorado Anschuz Medical Campus Health Data Compass; CUIMC: Columbia University Irving Medical Center; IQVIAHospitalCDM: IQVIA Hospital Charge Data Master; OHSU: Oregon Health and Science University; OPTUM-EHR: Optum© de-identified Electronic Health Record Dataset; OPTUM-SES: Optum® De-Identified Clinformatics® Data Mart Database – Socio-Economic Status (SES); STARR-OMOP: Stanford Medicine Research Data Repository; TRDW: Tufts MC Research Data Warehouse; UWM-CRD: UW Medicine COVID Research Dataset; VA-OMOP: Department of Veterans Affairs

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CHARYBDISKostkaDuarteSallesNatureSupplementaryMaterialsV3.pdf

- Tables.docx