Clinical Epidemiology

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Unraveling COVID-19: A Large-Scale Characterization of 4.5 Million COVID-19 Cases Using CHARYBDIS

Kristin Kostka[1,2], Talita Duarte-Salles[3], Albert Prats-Uribe[4], Anthony G Sena[5,6], Andrea Pistillo[3], Sara Khalid[4], Lana YH Lai[7], Asieh Golozar[8,9], Thamir M Alshammari[10], Dalia M Dawoud[11], Fredrik Nyberg[12], Adam B Wilcox[13,14], Alan Andryc[5], Andrew Williams[15], Anna Ostropolets[16], Carlos Areia[17], Chi Young Jung[18], Christopher A Harle[19], Christian G Reich[1,2], Clair Blacketer[5,6], Daniel R Morales[20], David A Dorr[21], Edward Burn[3,4], Elena Roel[3,22], Eng Hooi Tan[4], Evan Minty[23], Frank DeFalco[5], Gabriel de Maeztu[24], Gigi Lipori[19], Hiba Alghoul[25], Hong Zhu[26], Jason A Thomas[13], Jiang Bian[19], Jimyung Park[27], Jordi Martínez Roldán[28], Jose D Posada[29], Juan M Banda[30], Juan P Horcajada[31], Julianna Kohler[32], Karishma Shah[33], Karthik Natarajan[16,34], Kristine E Lynch[35,36], Li Liu[37], Lisa M Schilling[38], Martina Recalde[3,22], Matthew Spotnitz[14], Mengchun Gong[39], Michael E Matheny[40,41], Neus Valveny[42], Nicole G Weiskopf[21], Nigam Shah[29], Osaid Alser[43], Paula Casajust[42], Rae Woong Park[27,44], Robert Schuff[21], Sarah Seager[1], Scott L DuVall[35,36], Seng Chan You[45], Seokyoung Song[46], Sergio Fernández-Bertolín[3], Stephen Fortin[5], Tanja Magoc[19], Thomas Falconer[16], Vignesh Subbian[47], Vojtech Huser[48], Waheed-Ul-Rahman Ahmed[33,49], William Carter[38], Yin Guan[50], Yankuic Galvan[19], Xing He[19], Peter R Rijnbeek[6], George Hripcsak[16,34], Patrick B Ryan[5,16], Marc A Suchard[51], Daniel Prieto-Alhambra[4]

[1]IQVIA, Cambridge, MA, USA; [2]OHDSI Center at The Roux Institute, Northeastern University, Portland, ME, USA; [3]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain; [4]Centre for Statistics in Medicine, NDORMS, University of Oxford, Oxford, UK; [5]Janssen Research & Development, Titusville, NJ, USA; [6]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands; [7]School of Medical Sciences, University of Manchester, Manchester, UK; [8]Regeneron Pharmaceuticals, Tarrytown, NY, USA; [9]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA; [10]College of Pharmacy, Riyadh Elm University, Riyadh, Saudi Arabia; [11]National Institute for Health and Care Excellence, London, UK; [12]School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; [13]Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA; [14]Unviersity of Washington Medicine, Seattle, WA, USA; [15]Tufts Institute for Clinical Research and Health Policy Studies, Boston, MA, USA; [16]Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA; [17]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK; [18]Division of Respiratory and Critical Care Medicine, Department of Internal Medicine, Daegu Catholic University Medical Center, Daegu, South Korea; [19]University of Florida Health, Gainesville, FL, USA; [20]Division of Population Health and Genomics, University of Dundee, Dundee, UK; [21]Department of Medical Informatics & Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA; [22]Universitat Autònoma de Barcelona, Barcelona, Spain; [23]O'Brien Institute for Public Health, Faculty of Medicine, University of Calgary, Calgary, Canada; [24]IOMED, Barcelona, Spain; [25]Faculty of Medicine, Islamic University of Gaza, Gaza, Palestine; [26]Nanfang Hospital, Southern Medical University, Guangzhou, People's Republic of China; [27]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, South Korea; [28]Director of Innovation and Digital Transformation, Hospital del Mar, Barcelona, Spain; [29]Department of Medicine, School of Medicine, Stanford University, Redwood City, CA, USA; [30]Georgia State University, Department of Computer Science, Atlanta, GA, USA; [31]Department of Infectious Diseases, Hospital del Mar, Institut Hospital del Mar d'Investigació Mèdica (IMIM), Universitat Autònoma de Barcelona, Universitat Pompeu Fabra, Barcelona, Spain; [32]United States Agency for International Development, Washington, DC, USA; [33]Botnar Research Centre, NDORMS, University of Oxford, Oxford, UK; [34]New York-Presbyterian Hospital, New York, NY, USA; [35]VA Informatics and Computing Infrastructure, VA Salt Lake City Health Care System, Salt Lake City, UT, USA; [36]Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, USA; [37]Biomedical Big Data Center, Nanfang Hospital, Southern Medical University, Guangzhou, People's Republic of China; [38]Data Science to Patient Value Program, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA; [39]Institute of Health Management, Southern Medical University, Guangzhou, People's Republic of China; [40]Tennessee Valley Healthcare System, Veterans Affairs Medical Center, Nashville, TN, USA; [41]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, USA; [42]Real-World Evidence, TFS, Barcelona, Spain; [43]Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; [44]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea; [45]Department of Preventive Medicine, Yonsei University College of Medicine, Seoul, South Korea; [46]Department of Anesthesiology and Pain Medicine, Catholic University of Daegu, School of Medicine, Daegu, South Korea; [47]College of Engineering, The University of Arizona, Tucson, AZ, USA; [48]National Library of Medicine, National Institutes of Health, Bethesda, MD, USA; [49]College of Medicine and Health, University of Exeter, St Luke's Campus, Exeter, UK; [50]DHC Technologies Co. Ltd., Beijing, People's Republic of China; [51]Departments of Biostatistics, Computational Medicine, and Human Genetics, University of California, Los Angeles, CA, USA

Received: 18 June 2021
Accepted: 27 January 2022
Published: 22 March 2022

Correspondence: Daniel Prieto-Alhambra, Botnar Research Centre, Windmill Road, Oxford, OX37LD, UK, Email daniel.prietoalhambra@ndorms.ox.ac.uk

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

Patients and Methods: We conducted a descriptive retrospective database 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. 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*.

Results: We aggregated over 22,000 unique characteristics describing patients with COVID-19. All comorbidities, symptoms, medications, and outcomes are described by cohort in aggregate counts and are readily available online. Globally, we observed similarities in the USA and Europe: more women diagnosed than men but more men hospitalized than women, most diagnosed cases between 25 and 60 years of age versus most hospitalized cases between 60 and 80 years of age. South Korea differed with more women than men hospitalized. Common comorbidities included type 2 diabetes, hypertension, chronic kidney disease and heart disease. Common presenting symptoms were dyspnea, cough and fever. Symptom data availability was more common in hospitalized cohorts than diagnosed.

Conclusion: We constructed a global, multi-centre view to describe trends in COVID-19 progression, management and evolution over time. By characterising baseline variability in patients and geography, our work provides critical context that may otherwise be misconstrued as data quality issues. This is important as we perform studies on adverse events of special interest in COVID-19 vaccine surveillance.

Keywords: OHDSI, OMOP CDM, descriptive epidemiology, real world data, real world evidence, open science

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 the end of 2021, the number of COVID-19 cases increased to over 278 million cases globally, and the death toll exceeded 5 million.[6] Thousands of publications have attempted to aid our scientific understanding of this public health emergency.[7,8]

Characterisation studies, called descriptive epidemiology, provide an important context into our understanding of disease by describing the basic attributes of who gets sick and in what context. The initial body of COVID-19 characterisation work gave researchers information on the stark difference in the perception of the novel coronavirus compared to flu-like illnesses: patients were male, younger, and with fewer concurrent comorbidities and less documented prior medication use.[9]

Utilising routinely collected real world data (RWD) can be a powerful asset for understanding 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,[10] 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.[11,12] Here we present the global Observational Health Data Sciences and Informatics (OHDSI) community, an international open-science initiative of more than 3500 collaborators from 34 countries, 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).[9] 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. Herein we aim to 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. These data reflect an international community of research collaborators who are working to advance retrospective database research in RWD for COVID-19. Our body of research is freely available, foundational result set that can provide benchmarks in how COVID-19 manifests over time including its inevitable evolution as we roll-out additional vaccines and treatments.

Methods

Study Design, Setting and Data Sources

We conducted a descriptive retrospective database 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). Each data partner mapped 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 assessed the plausibility, conformance and completeness of each contributing database using a common data quality tool for repeated assessment and monitoring the adherence to conventions across the network.[16,17] We ensured technical 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] 23 real world healthcare databases contributed to the CHARYBDIS study (Supplementary Table 1). Contributing institutes ranged from major academic medical centers to small community hospitals from across three continents. Date capture ranged from December 2019 to as recent as January 2021 (site specific dates in Supplementary Table 1). 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 code documenting COVID-19 - earliest event served as the index date); ii) *hospitalized with COVID-19* and; iii) *hospitalized with COVID-19 and requiring intensive services*. Due to variability in access to diagnostic testing, we specifically looked for the presence of a PCR or antigen laboratory test OR the use of clinical diagnosis codes documenting COVID-19 presentation.[19] 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. Databases 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.[20] 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 the explicit intention of building an initial, repeatable framework for constructing prevalent rates of disease. 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.

Results

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 and 4).

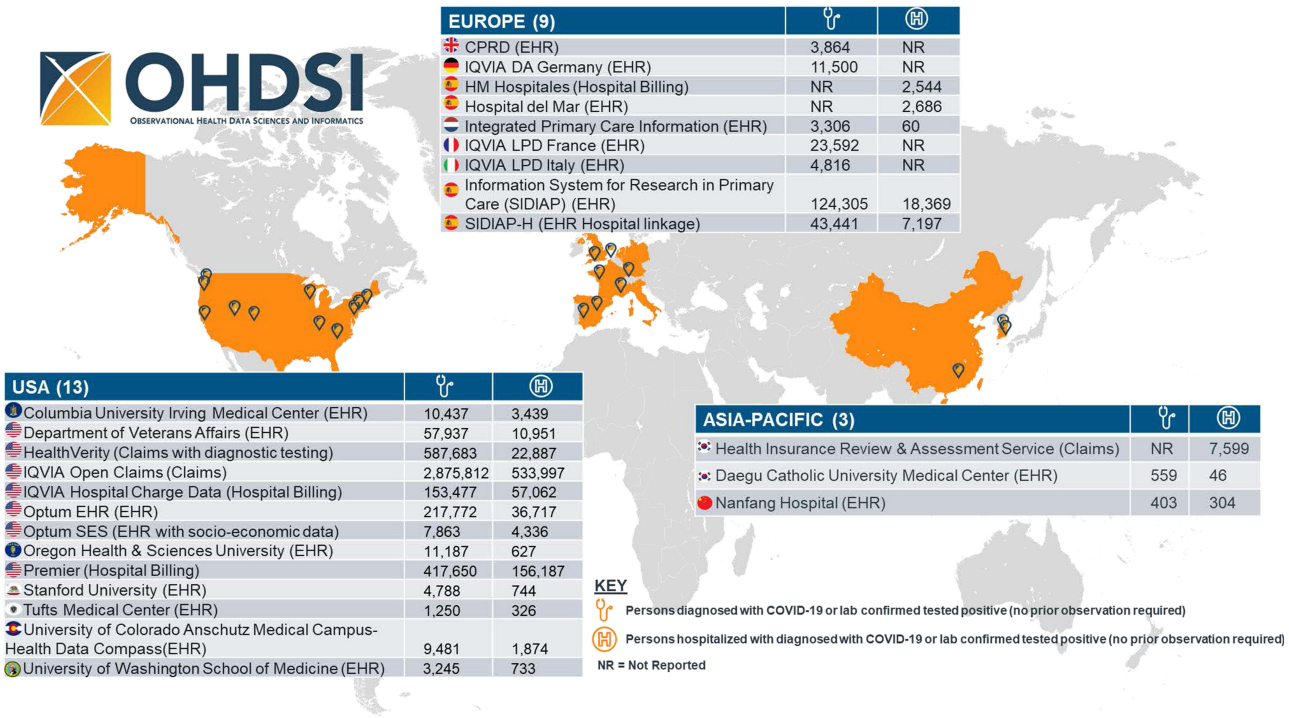

Figure 1 COVID-19 cases across the OHDSI COVID-19 network.

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 hospitalized 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

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.

Notes: 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; CU-AMC-HDC, U of Colorado Anschutz 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.

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 hospitalized 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 hospitalized 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 hospitalized cohort, the proportion of which was < 2%.

Other Analyses

Dyspnea, cough, and fever were the most common symptoms in diagnosed and hospitalized cohorts globally (Supplementary Table 5). Where recorded, the proportion of dyspnea and malaise/fatigue was consistently higher in the hospitalized 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 hospitalized cohorts (Supplementary Table 6).

We further described a total of 19,222 conditions and 2973 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 (Supplementary Tables 4–6) The full result set of 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/.

Discussion

CHARYBDIS is the world's largest open science aggregate result set aimed at describing the baseline demographics, clinical characteristics, treatments received, and outcomes among individuals diagnosed and hospitalized with COVID-19. To accomplish this, we aggregated over 22,000 unique characteristics creating a multi-centre view to describe trends in COVID-19 progression, management and evolution over time. Globally, we observed similarities in the USA and Europe in gender (more women diagnosed than men but more men hospitalized than women) and age (most diagnosed cases between 25–60 years of age versus most hospitalized cases between 60–80 years of age) distributions. Similar to previous studies, we observed South Korea differed with more women than men hospitalized. We found similarities in comorbidities and presenting symptoms. The large, diverse sample size allows also for the identification of populations of great interest, including children and adolescents,[25] pregnant women,[26] patients with a history of cancer,[27] patients with a history of autoimmune disorders,[28] or patterns of drug utilization in COVID-19 treatment,[21] and which were the focus of additional in-depth investigations.

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 evidence framework is a method for systematically understanding cohort-level differences in COVID-19 from different regions and different points in the pandemic. In the months since we started this effort, our network has already aided in rapid study for coagulopathy and adverse of events of special interest for COVID-19 vaccines to inform regulatory bodies.[22] This research community can be a public health utility to guide in 1) better patient characterization and stratification, 2) identifying areas of gap in knowledge/evidence, and 3) generating hypotheses for future research.

Table 1 Characteristics of Persons with a COVID-19 Diagnosis or SARS-CoV-2 Positive Test Across the OHDSI COVID-19 Network*

	Asia		United States													Europe					
	DCMC	NFHCRD	Health Verity	Premier	OPTUM-EHR	OPTUM-SES	STARR-OMOP	TRDW	VA-OMOP	IQVIA-OpenClaims	IQVIA Hospital CDM	CUIMC	CU-AMC-HDC	UWM-CRD	OHSU	SIDIAP	IPCI	CPRD	IQVIA LPD France	IQVIA DA Germany	IQVIA LPD Italy
COVID-19 Cases (N)	559	403	587,683	66,132	160,613	7863	4788	1250	57,937	2,785,812	153,477	10,437	9481	3245	11,187	124,305	3306	3864	23,592	11,500	4816
Persons Tested	NR	397	3,898,593	219,230	1,025,584	41,673	56,881	6950	521,814	6,520,151	719,596	22,094	120,661	83,921	109,434	173,957	NR	5551	NR	NR	NR
Tested Positive, n (%)*	NR	392 (97.3)	425,610 (72.4)	NR	73,113 (45.5)	NR	1880 (39.3)	1035 (82.8)	32,847 (56.7)	NR	NR	6959 (66.7)	NR	3,140 (96.8)	8764 (78.3)	39,047 (31.4)	NR	2098 (54.3)	NR	NR	NR
Full 30-day follow up	162 (29.0)	276 (68.5)	67,071 (11.4)	3902 (5.9)	84,073 (52.3)	1269 (16.1)	2703 (56.5)	641 (51.3)	44,661 (77.1)	1,882,950 (67.6)	21,145 (13.8)	2008 (19.2)	8755 (92.3)	1,199 (36.9)	3760 (33.6)	81,914 (65.9)	2601 (78.7)	2723 (70.5)	9819 (41.6)	5588 (48.6)	3570 (74.1)
< 30-day follow up	397 (71.0)	127 (31.5)	520,612 (88.6)	62,230 (94.1)	76,540 (47.7)	6594 (83.9)	2085 (43.5)	609 (48.7)	13,272 (22.9)	902,862 (32.4)	132,332 (86.2)	8429 (80.8)	706 (7.4)	2046 (63.1)	7427 (66.4)	42,391 (34.1)	705 (21.3)	1141 (29.5)	13,773 (58.4)	5912 (51.4)	1246 (25.9)
Comorbidities, n (%)**																					
Type 2 Diabetes Mellitus	108 (19.3)	9 (2.2)	20,922 (3.6)	10,783 (16.3)	26,897 (16.7)	2673 (34.0)	555 (11.6)	179 (14.3)	19,083 (32.9)	724,991 (26.0)	35,576 (23.2)	1977 (18.9)	1396 (14.7)	391 (12.0)	603 (5.4)	9941 (8.0)	500 (15.1)	545 (14.1)	1318 (5.6)	1089 (9.5)	452 (9.4)
Hypertension	154 (27.5)	19 (4.7)	34,090 (5.8)	19,008 (28.7)	54,678 (34.0)	4393 (55.9)	1319 (27.5)	307 (24.6)	34,357 (59.3)	1,260,816 (45.3)	60,495 (39.4)	3771 (36.1)	2708 (28.6)	735 (22.7)	1065 (9.5)	21,337 (17.2)	688 (20.8)	779 (20.2)	3522 (14.9)	2611 (22.7)	1659 (34.4)
Heart disease	106 (19.0)	7 (1.7)	19,016 (3.2)	11,533 (17.4)	39,510 (24.6)	3726 (47.4)	977 (20.4)	245 (19.6)	24,699 (42.6)	936,271 (33.6)	33,846 (22.1)	3236 (31.0)	1871 (19.7)	440 (13.6)	778 (7.0)	17,759 (14.3)	470 (14.2)	722 (18.7)	1213 (5.1)	2007 (17.5)	1013 (21.0)
History of cancer	32 (5.7)	NR	6107 (1.0)	3157 (4.8)	18,536 (11.5)	1491 (19.0)	887 (18.5)	106 (8.5)	10,792 (18.6)	317,479 (11.4)	11,237 (7.3)	1480 (14.2)	843 (8.9)	184 (5.7)	469 (4.2)	8872 (7.1)	262 (7.9)	296 (7.7)	674 (2.9)	661 (5.7)	547 (11.4)
Hepatitis C	NR	NR	740 (0.1)	410 (0.6)	1395 (0.9)	112 (1.4)	61 (1.3)	35 (2.8)	3075 (5.3)	40,101 (1.4)	1966 (1.3)	144 (1.4)	90 (0.9)	54 (1.7)	88 (0.8)	648 (0.5)	NR	NR	40 (0.2)	31 (0.3)	53 (1.1)
Obesity	29 (5.2)	NR	15,072 (2.6)	7298 (11.0)	71,076 (44.3)	2468 (31.4)	1246 (26.0)	325 (26.0)	25,128 (43.4)	740,430 (26.6)	28,757 (18.7)	3729 (35.7)	3136 (33.1)	233 (7.2)	945 (8.4)	36,557 (29.4)	629 (19.0)	1428 (37.0)	2287 (9.7)	1345 (11.7)	674 (14.0)
Dementia	6 (1.1)	NR	4255 (0.7)	3697 (5.6)	5360 (3.3)	851 (10.8)	38 (0.8)	29 (2.3)	4019 (6.9)	219,062 (7.9)	7776 (5.1)	483 (4.6)	235 (2.5)	116 (3.6)	97 (0.9)	6013 (4.8)	64 (1.9)	327 (8.5)	55 (0.2)	339 (2.9)	81 (1.7)
Autoimmune condition	49 (8.8)	NR	7291 (1.2)	1678 (2.5)	13,396 (8.3)	1464 (18.6)	418 (8.7)	133 (10.6)	10,103 (17.4)	433,259 (15.6)	8965 (5.8)	1388 (13.3)	720 (7.6)	140 (4.3)	409 (3.7)	8260 (6.6)	476 (14.4)	394 (10.2)	1467 (6.2)	1183 (10.3)	636 (13.2)
Chronic obstructive pulmonary disease (COPD) without asthma	NR	NR	8160 (1.4)	3335 (5.0)	12,067 (7.5)	1449 (18.4)	231 (4.8)	89 (7.1)	12,665 (21.9)	297,269 (10.7)	12,008 (7.8)	809 (7.8)	733 (7.7)	112 (3.5)	249 (2.2)	15,819 (12.7)	213 (6.4)	294 (7.6)	696 (3.0)	868 (7.5)	350 (7.3)
Asthma without COPD	17 (3.0)	NR	10,458 (1.8)	3972 (6.0)	21,076 (13.1)	1125 (14.3)	521 (10.9)	112 (9.0)	6278 (10.8)	438,892 (15.8)	12,936 (8.4)	1376 (13.2)	1100 (11.6)	176 (5.4)	567 (5.1)	7567 (6.1)	322 (9.7)	494 (12.8)	2327 (9.9)	1097 (9.5)	420 (8.7)
Pregnant women	NR	NR	3543 (0.6)	1192 (1.8)	3917 (2.4)	109 (1.4)	52 (1.1)	27 (2.2)	86 (0.1)	41,329 (1.5)	2944 (1.9)	382 (3.7)	212 (2.2)	32 (1.0)	156 (1.4)	689 (0.6)	32 (1.0)	11 (0.3)	212 (0.9)	39 (0.3)	63 (1.3)

(*Continued*)

https://doi.org/10.2147/CLEP.S323292

DovePress

Table 1 (Continued).

	Asia		United States													Europe					
	DCMC	NFHCRD	Health Verity	Premier	OPTUM-EHR	OPTUM-SES	STARR-OMOP	TRDW	VA-OMOP	IQVIA-OpenClaims	IQVIA Hospital CDM	CUIMC	CU-AMC-HDC	UWM-CRD	OHSU	SIDIAP	IPCI	CPRD	IQVIA LPD France	IQVIA DA Germany	IQVIA LPD Italy
Chronic kidney disease broad	156 (27.9)	NR	7535 (1.3)	5711 (8.6)	17,531 (10.9)	1829 (23.3)	398 (8.3)	NR	10,239 (17.7)	364,857 (13.1)	16,250 (10.6)	1181 (11.3)	723 (7.6)	213 (6.6)	277 (2.5)	8144 (6.6)	197 (6.0)	478 (12.4)	194 (0.8)	562 (4.9)	192 (4.0)
End stage renal disease	155 (27.7)	NR	1683 (0.3)	1062 (1.6)	3008 (1.9)	359 (4.6)	122 (2.5)	NR	3273 (5.6)	96,555 (3.5)	5155 (3.4)	600 (5.7)	166 (1.8)	51 (1.6)	52 (0.5)	8 (0.0)	NR	17 (0.4)	NR	27 (0.2)	NR
Human immunodeficiency virus infection	NR	NR	829 (0.1)	357 (0.5)	763 (0.5)	67 (0.9)	20 (0.4)	NR	817 (1.4)	24,808 (0.9)	1309 (0.9)	163 (1.6)	56 (0.6)	45 (1.4)	43 (0.4)	290 (0.2)	NR	NR	83 (0.4)	18 (0.2)	19 (0.4)

Notes: *Proportions presented among diagnosed patients with a COVID-19 diagnosis or SARS-CoV-2 positive test by database (column percentage); since SIDIAP_H includes a subset of SIDIAP, results were not included in this table; - data not available or below the minimum cell count required (5 individuals); no prior observation time was required. **Prevalent conditions at index date.

Abbreviations: CU-AMC-HDC, U of Colorado Anschutz 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 Research Data Warehouse; UWM-CRD, UW Medicine COVID Research Dataset; VA-OMOP, Department of Veterans Affairs; NR, not reported by data partner.

Table 2 Characteristics of Persons Hospitalized with a COVID-19 Diagnosis or SARS-CoV-2 Positive Test Across the OHDSI COVID-19 Network*

	Asia		United States													Europe		
	HIRA	**NFHCRD**	**HealthVerity**	**Premier**	**OPTUM-EHR**	**OPTUM-SES**	**STARR-OMOP**	**TRDW**	**VA-OMOP**	**IQVIA Open Claims**	**IQVIA Hospital CDM**	**CUIMC**	**CU-AMC-HDC**	**UWM-CRD**	**OHSU**	**HM Hospitales**	**SIDIAP**	**HMAR**
COVID-19 Cases (N)	7599	304	22,887	36,019	29,061	4336	744	326	10,951	533,997	57,062	3439	1874	733	627	2544	18,369	2686
Hospitalized with positive test, n (%)	NR	125 (41.1)	13,262 (57.9)	NR	13,817 (47.5)	NR	128 (17.2)	232 (71.2)	8623 (78.7)	NR	NR	3075 (89.4)	NR	676 (92.2)	344 (54.9)	NR	13,685 (74.5)	773 (28.8)
Full 30-day follow up	7359 (96.8)	284 (93.4)	10,333 (45.1)	2361 (6.6)	18,555 (63.8)	851 (19.6)	657 (88.3)	NR	8548 (78.1)	412,537 (77.3)	11,876 (20.8)	943 (27.4)	1810 (96.6)	400 (54.6)	484 (77.2)	109 (4.3)	12,290 (66.9)	1254 (46.7)
< 30-day follow up	240 (3.2)	20 (6.6)	12,554 (54.9)	33,658 (93.4)	10,506 (36.2)	3485 (80.4)	87 (11.7)	NR	2400 (21.9)	121,460 (22.7)	45,186 (79.2)	2496 (72.6)	64 (3.4)	333 (45.4)	143 (22.8)	2435 (95.7)	6079 (33.1)	1432 (53.3)
Comorbidities, n (%)**																		
Type 2 Diabetes Mellitus	1760 (23.2)	NR	3880 (17.0)	8899 (24.7)	9531 (32.8)	1844 (42.5)	157 (21.1)	83 (25.5)	5839 (53.3)	254,505 (47.7)	16,480 (28.9)	1120 (32.6)	677 (36.1)	226 (30.8)	177 (28.2)	428 (16.8)	3295 (17.9)	294 (10.9)
Hypertension	1943 (25.6)	NR	6410 (28.0)	15,216 (42.2)	16,427 (56.5)	2977 (68.7)	389 (52.3)	123 (37.7)	9087 (83.0)	390,171 (73.1)	26,262 (46.0)	1770 (51.5)	1073 (57.3)	398 (54.3)	283 (45.1)	1139 (44.8)	5645 (30.7)	653 (24.3)
Heart disease	1271 (16.7)	NR	5178 (22.6)	10,384 (28.8)	13,274 (45.7)	2634 (60.7)	297 (39.9)	109 (33.4)	7421 (67.8)	319,842 (59.9)	16,165 (28.3)	1534 (44.6)	802 (42.8)	286 (39.0)	258 (41.1)	606 (23.8)	5148 (28.0)	362 (13.5)
History of cancer	410 (5.4)	NR	1132 (4.9)	2811 (7.8)	4939 (17.0)	1065 (24.6)	277 (37.2)	47 (14.4)	3401 (31.1)	106,805 (20.0)	5524 (9.7)	588 (17.1)	300 (16.0)	90 (12.3)	154 (24.6)	286 (11.2)	2616 (14.2)	179 (6.7)
Hepatitis C	61 (0.8)	NR	134 (0.6)	394 (1.1)	469 (1.6)	77 (1.8)	17 (2.3)	13 (4.0)	1037 (9.5)	14,408 (2.7)	1050 (1.8)	81 (2.4)	37 (2.0)	25 (3.4)	37 (5.9)	15 (0.6)	135 (0.7)	38 (1.4)
Obesity	16 (0.2)	NR	2238 (9.8)	6678 (18.5)	15,497 (53.3)	1626 (37.5)	312 (41.9)	126 (38.7)	5677 (51.8)	191,071 (35.8)	10,735 (18.8)	1651 (48.0)	988 (52.7)	138 (18.8)	167 (26.6)	149 (5.9)	8428 (45.9)	259 (9.6)
Dementia	436 (5.7)	NR	1815 (7.9)	3428 (9.5)	2376 (8.2)	637 (14.7)	17 (2.3)	17 (5.2)	2087 (19.1)	81,638 (15.3)	4044 (7.1)	373 (10.8)	140 (7.5)	95 (13.0)	25 (4.0)	108 (4.2)	1102 (6.0)	75 (2.8)
Autoimmune condition	813 (10.7)	NR	1215 (5.3)	1432 (4.0)	3320 (11.4)	931 (21.5)	89 (12.0)	54 (16.6)	3156 (28.8)	136,735 (25.6)	4205 (7.4)	570 (16.6)	226 (12.1)	67 (9.1)	83 (13.2)	121 (4.8)	1706 (9.3)	93 (3.5)
Chronic obstructive pulmonary disease (COPD) without asthma	145 (1.9)	NR	2213 (9.7)	3016 (8.4)	5176 (17.8)	1066 (24.6)	102 (13.7)	52 (16.0)	4641 (42.4)	118,421 (22.2)	7071 (12.4)	469 (13.6)	333 (17.8)	77 (10.5)	90 (14.4)	173 (6.8)	4848 (26.4)	138 (5.1)
Asthma without COPD	1560 (20.5)	NR	1004 (4.4)	2677 (7.4)	3746 (12.9)	628 (14.5)	127 (17.1)	39 (12.0)	1153 (10.5)	82,087 (15.4)	3825 (6.7)	498 (14.5)	245 (13.1)	58 (7.9)	101 (16.1)	112 (4.4)	957 (5.2)	99 (3.7)
Pregnant women	121 (1.6)	NR	341 (1.5)	682 (1.9)	1550 (5.3)	30 (0.7)	18 (2.4)	13 (4.0)	NR	12,748 (2.4)	2029 (3.6)	158 (4.6)	111 (5.9)	22 (3.0)	73 (11.6)	7 (0.3)	108 (0.6)	20 (0.7)
Chronic kidney disease broad	421 (5.5)	NR	2622 (11.5)	5339 (14.8)	6596 (22.7)	1357 (31.3)	162 (21.8)	NR	3958 (36.1)	164,710 (30.8)	8827 (15.5)	691 (20.1)	375 (20.0)	152 (20.7)	112 (17.9)	157 (6.2)	2658 (14.5)	186 (6.9)

(*Continued*)

https://doi.org/10.2147/CLEP.S323292 DovePress

Table 2 (Continued).

	Asia		United States													Europe		
	HIRA	**NFHCRD**	**HealthVerity**	**Premier**	**OPTUM-EHR**	**OPTUM-SES**	**STARR-OMOP**	**TRDW**	**VA-OMOP**	**IQVIA Open Claims**	**IQVIA Hospital CDM**	**CUIMC**	**CU-AMC-HDC**	**UWM-CRD**	**OHSU**	**HM Hospitales**	**SIDIAP**	**HMAR**
End stage renal disease	30 (0.4)	NR	826 (3.6)	948 (2.6)	1506 (5.2)	296 (6.8)	31 (4.2)	NR	1520 (13.9)	53,747 (10.1)	3333 (5.8)	371 (10.8)	101 (5.4)	43 (5.9)	29 (4.6)	7 (0.3)	NR	91 (3.4)
Human immuno-deficiency virus infection	NR	NR	96 (0.4)	275 (0.8)	222 (0.8)	33 (0.8)	NR	NR	239 (2.2)	7009 (1.3)	516 (0.9)	73 (2.1)	14 (0.7)	11 (1.5)	11 (1.8)	NR	47 (0.3)	14 (0.5)

Notes: *Proportions presented among diagnosed patients with a COVID-19 diagnosis or SARS-CoV-2 positive test by database (column percentage); - data not available or below the minimum cell count required (5 individuals); no prior observation time was required. **Prevalent conditions at index date.

Abbreviations: CU-AMC-HDC, U of Colorado Anschutz 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 Research Data Warehouse; UWM-CRD, UW Medicine COVID Research Dataset; VA-OMOP, Department of Veterans Affairs; HM-Hospitales, HM-Hospitales Madrid; SIDIAP, Information System for Research in Primary Care; HMAR, Hospital del Mar; NR, not reported by data partner.

Comparison to Other Multi-Centre COVID-19 Consortia

We began our deep phenotyping work through an initial investigation of persons hospitalized with COVID-19 compared to prior flu seasons in our global federated network.[9]

The National COVID Cohort Collaborative (N3C) is a NIH NCATS funded initiative collecting centralizing patient-level data to study patterns in COVID-19 patients.[23] This effort has over 80 participating institutions contributing 4.5M COVID-19 patients to date to a centralized harmonized repository. The consortia has enabled many US institutions in adoption of common data models in COVID-19 research. 4CE is another multi-site data-sharing collaborative of 342 hospitals in the US and in Europe, utilizing i2b2 or OMOP data models.[24] The hospitalization cohorts presented in 4CE cohorts remain smaller than the scope of CHARYBDIS with only 36,447 hospitalized patients with COVID-19 as of August 2020.[24] Even when adjusting for cohort overlap, our work to date with CHARYBDIS is nearly triple the diagnosis and double the hospitalized cohorts represented in prior research. Our results also have more international representation across the cascade of hotspots over the course of the pandemic's spread. As we continue our research, we are working with researchers to create 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 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 for extensive investigation on subgroups of interest. CHARYBDIS is the framework for additional in-depth investigations on children and adolescents,[25] pregnant women,[26] patients with a history of cancer,[27] patients with a history of autoimmune disorders,[21] or patterns of drug utilization in COVID-19 treatment.[21] The size of these results are so large, we have hundreds combinations of subgroups of interest that remain unreported. There is significant opportunity for this framework to inform additional research.

Study Limitations

We recognize there are limitations in our approach. First, this study is descriptive in nature. Further analyses are needed to utilize these findings in clinical application. The observed differences between groups (eg diagnosed versus hospitalized) should therefore not be interpreted as causal effects without further statistical scrutiny. 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. We are aware that this may mean the network assembled is not inherently valuable for every follow-on analysis as each data partner may have different elements missing. Hospital encounters may be unable to ascertain outcomes experienced in an outpatient data. Our EHR partners rely on structured data and may be missing key findings from clinical notes. 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. Further analysis using this network requires

stratification by calendar month. Lastly, we currently lack data partners in low to middle income countries and recognize these data are lacking representation of some of the hardest hit areas in the world (eg Brazil, India). 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 pandemic.

Conclusion

We constructed a global, multi-centre view to describe trends in COVID-19 progression, management and evolution over time. By characterising baseline variability in demographics across geography, our work provides critical context to the reliability of the insight we generate. In retrospective database studies, one can struggle to identify whether heterogeneity occurs because of patient variability or because of the variability in source systems we use to capture patient data. Here we use a network of retrospective databases standardised to the same data model adhering to a shared ontology and data quality processes. Our study provides a comprehensive view into the first year of the pandemic at a scale unlike most retrospective research. Our work sheds light on the natural history of millions of COVID-19 patients from the USA, 6 European countries and 2 Asian countries. This framework is open source and available for re-use enabling a repeatable, reproducible method to capture the evolving natural history of this novel coronavirus and can be extended to other disease of international interest. We believe it is critically important to repeat and reproduce the findings we produce in real world studies. Leveraging this global federated network to corroborate single center findings can provide context to national database findings in the presence of regional variability in COVID-19 management including vaccine rollout and treatments.

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.

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 (eg 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.

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.

Acknowledgments

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. We also thank the database curation teams around the world including the COVIDMAR Group (R.Güerri, J.Villar, L.Sorlí, M.Montero, S.Gómez-Zorrilla, I. López-Montesinos, M.Arenas-Miras, J.Gómez-Junyent, I.Arrieta, E.Sendra, S.Castañeda, E.Letang, I.Pelegrín, A.Rial, J. Rodríguez, C.Gimenez, J.Soldado, E.García). Kristin Kostka and Talita Duarte-Salles are co-first authors for this study. Marc A Suchard and Daniel Prieto-Alhambra are co-senior authors for this study.

Author Contributions

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, PRR, 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, JMR, JPH, 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. All authors made a significant contribution to the work reported, whether that is in the conception, study design, execution, acquisition of data, analysis and interpretation, or in all these areas; took part in drafting, revising or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

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, IQVIA, TFS and IOMED. 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). TFS received funding related to this work from the University of Oxford. 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. The Ajou University received funding related to this work from the Bill & Melinda Gates Foundation (Investment ID INV-016284), from the Bio Industrial Strategic Technology Development Program (20003883), funded by the Ministry of Trade, Industry & Energy, and from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, funded by the Ministry of Health & Welfare, Republic of Korea (HR16C0001).

Disclosure

Ms. Kostka was an employee of IQVIA during the conduct of this study and received grant funding from the NIH NCATS National COVID Cohort Collaborative and the Bill and Melinda Gates Foundation. 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 was an employee of AstraZeneca until 2019 and hold some shares. 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. Mr. Thomas reports grants from Bill and Melinda Gates Foundation (INV-016910), grants from National Center for Advancing Translational Sciences (NCATS), National Institutes of Health, through Grant Award Number UL1TR002369 to his institution, during the conduct of the study. Dr Jiang Bian reports grants from NIH/NIEHS (R21ES032762), 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, during the conduct of the study and 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. Mr Robert Schuff reports grants from Gates Foundation, grants from NIH-NCATS, during the conduct of the study. Ms. Seager is an employee of IQVIA. Dr. DuVall reports grants from Anolinx, LLC, Astellas Pharma, Inc, AstraZeneca Pharmaceuticals LP, Boehringer Ingelheim International GmbH, Celgene Corporation, Eli Lilly and Company, Genentech Inc., Genomic Health, Inc., Gilead Sciences Inc., GlaxoSmithKline PLC, Innocrin Pharmaceuticals Inc., Janssen Pharmaceuticals, Inc., Kantar Health, Myriad Genetic Laboratories, Inc., Novartis International AG, and 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. Subbian reports grants from State of Arizona; Arizona Board of Regents, during the conduct of the study; grants from National Science Foundation (grant# 1838745), 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, Janssen Research and Development, during the conduct of the study. He also works for a research institute which receives/received unconditional research grants from

Yamanouchi, Pfizer-Boehringer Ingelheim, GSK, Amgen, UCB, Novartis, Astra-Zeneca, Chiesi, Janssen Research and Development, none of which relate to the content of this work. Dr. Hripcsak reports grants from US NIH and Janssen Research. Dr. Ryan is an employee of Janssen Research and Development and shareholder of Johnson & Johnson. Dr. Suchard reports grants from US National Institutes of Health, Department of Veterans Affairs, during the conduct of the study; grants and/or personal fees from IQVIA, Janssen Research and Development, US Food and Drug Administration, and Private Health Management, outside the submitted work. Dr. Prieto-Alhambra reports grants, non-financial support, speaker/consultancy services and/or advisory board membership from AMGEN, UCB Biopharma, and 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. The authors report no other conflicts of interest in this work.

References

1. Kent S, Burn E, Dawoud D, et al. Common problems, common data model solutions: evidence generation for health technology assessment. *Pharmacoeconomics*. 2020;39:275–285. doi:10.1007/s40273-020-00981-9
2. Forrest CB, McTigue KM, Hernandez AF, 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, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform*. 2015;216:574–578.
4. Sena A, Kostka K, Schuemie M, Posada JD, Schuemie M. 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; 2021. Available from: 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. Johns Hopkins Coronavirus Resource Center. COVID-19 map; 2021. Available from: https://coronavirus.jhu.edu/map.html. Accessed March 4, 2022.
7. COVID-19-related medical research: a meta-research and critical appraisal; 2021. Available from: 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. Burn E, You SC, Sena AG, et al. Deep phenotyping of 34,128 adult patients hospitalized with COVID-19 in an international network study. *Nat Commun*. 2020;11:5009. doi:10.1038/s41467-020-18849-z
10. Subbian V, Solomonides A, Clarkson M, 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;27. doi:10.1093/jamia/ocaa188
11. Madhavan S, Bastarache L, Brown JS, 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
12. Williamson EJ, Walker AJ, Bhaskaran K, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. 2020;584:430–436. doi:10.1038/s41586-020-2521-4
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, 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, 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; 2021. Available from: https://ohdsi.github.io/TheBookOfOhdsi/DataQuality.html#data-quality-in-general. Accessed March 4, 2022.
18. Schuemie MJ, Cepeda MS, Suchard MA, 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. Kadri SS, Gundrum J, Warner S, et al. Uptake and accuracy of the diagnosis code for COVID-19 among US hospitalizations. *JAMA*. 2020;324:2553–2554. doi:10.1001/jama.2020.20323
20. HADES. Observational health data sciences and informatics; 2021. Available from: https://ohdsi.github.io/Hades/index.html. Accessed March 4, 2022.
21. Prats-Uribe A, Sena AG, Lai LYH, et al. Use of repurposed and adjuvant drugs in hospital patients with covid-19: multinational network cohort study. *BMJ*. 2021;373:n1038. doi:10.1136/bmj.n1038
22. Li X, Ostropolets A, Makadia R, Shoaibi A, Rao G, Sena AG et al. Characterising the background incidence rates of adverse events of special interest for covid-19 vaccines in eight countries: multinational network cohort study. *BMJ*. 2021; 373 :n1435. doi:10.1136/bmj.n1435
23. 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

24. Weber GM, Hong C, Palmer NP, et al.; 4CE Collaborative. 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
25. Duarte-Salles T, Vizcaya D, Pistillo A, et al. Thirty-Day Outcomes of Children and Adolescents With COVID-19: An International Experience. *Pediatrics September*. 2021; 148 (3): e2020042929. doi:10.1542/peds.2020-042929
26. Lai LYH, Golozar A, Sena A, et al. Clinical characteristics, symptoms, management and health outcomes in 8598 pregnant women diagnosed with COVID-19 compared to 27,510 with seasonal influenza in France, Spain and the US: a network cohort analysis. *medRxiv*. 2020. doi:10.1101/2020.10.29.20222083
27. Roel E, Pistillo A, Recalde M, et al. Characteristics and Outcomes of Over 300,000 Patients with COVID-19 and History of Cancer in the United States and Spain. *Cancer Epidemiol Biomarkers Prev*. 1 October 2021; 30 (10): 1884–1894.
28. Tan EH, Sena AG, Prats-Uribe A, et al. COVID-19 in patients with autoimmune diseases: characteristics and outcomes in a multinational network of cohorts across three countries. *Rheumatology*. 2021;60:SI37–SI50. doi:10.1093/rheumatology/keab250