TITLE: Baseline phenotype and 30-day outcomes of people tested for COVID-19: an international network cohort including >3.32 million people tested with real-time PCR and >219,000 tested positive for SARS-CoV-2 in South Korea, Spain and the United States

AUTHORS: Asieh Golozar[1, 2]* †, Lana YH Lai[3]*, Anthony G. Sena[4, 5], David Vizcaya[6], Lisa M. Schilling[7], Vojtech Huser[8], Fredrik Nyberg[9], Scott L. Duvall[10], Daniel R. Morales[11], Thamir M Alshammari[12], Hamed Abedtash[13], Waheed-Ul-Rahman Ahmed[14,15], Osaid Alser[16], Heba Alghoul[17], Ying Zhang[18], Mengchun Gong[18], Yin Guan[18], Carlos Areia[19], Jitendra Jonnagaddala[20], Karishma Shah[14], Jennifer C.E. Lane[14], Albert Prats-Uribe[14], Jose D. Posada[21], Nigam H. Shah[21], Vignesh Subbian[22], Lin Zhang[23, 24], Maria Tereza Fernandes Abrahão[24], Peter R. Rijnbeek[5], Seng Chan You[26], Paula Casajust[27], Elena Roel[28], Martina Recalde[28, 29], Sergio Fernández-Bertolín[28], Alan Andryc[4], Jason A. Thomas[30], Adam B. Wilcox[30, 31], Stephen Fortin[32], Clair Blacketer[4, 5], Frank DeFalco[4], Karthik Natarajan[33, 34], Thomas Falconer[33], Matthew Spotnitz[33], Anna Ostropolets[33], George Hripcsak[33, 34], Marc Suchard[35], Kristine E. Lynch[10], Michael E. Matheny[36], Andrew Williams[37], Christian Reich[38], Talita Duarte-Salles[28], Kristin Kostka[38], Patrick B. Ryan[4,5], and Daniel Prieto-Alhambra[39]

* Joint first author
† Corresponding author

AFFILIATIONS:

1. Regeneron Pharmaceutical, NY USA
2. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, MD USA
3. Division of Cancer Sciences, School of Medical Sciences, University of Manchester, UK
4. Janssen R&D, Titusville NJ, USA
5. Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
6. Bayer Pharmaceuticals, Sant Joan Despi, Spain
7. Data Science to Patient Value Program, University of Colorado Anschutz Medical Campus
8. National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
9. School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden
10. VINCI, VA Salt Lake City Health Care System, Salt Lake City, VA, & Division of Epidemiology, University of Utah, Salt Lake City, UT
11. Division of Population Health and Genomics, University of Dundee, UK
12. Medication Safety Research Chair, King Saud University, Riyadh, Saudi Arabia
13. Eli Lilly and Company, Lilly Corporate Center, Indianapolis, IN
14. Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
15. College of Medicine and Health, University of Exeter, St Luke's Campus, Heavitree Road, Exeter, EX1 2LU, UK.
16. Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
17. Faculty of Medicine, Islamic University of Gaza, Palestine
18. DHC Technologies Co. Ltd, Beijing, China
19. Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK
20. School of Public Health and Community Medicine, UNSW Sydney, Australia
21. Stanford University School of Medicine, Stanford, California, USA
22. College of Engineering, The University of Arizona, Tucson, Arizona, USA
23. School of Population Medicine and Public Health, Peking Union Medical College and Chinese Academy of Medical Sciences, Beijing, China

24. Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia
25. Faculty Medicine University of São Paulo, São Paulo, Brazil
26. Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, South Korea
27. Real-World Evidence, Trial Form Support, Barcelona, Spain
28. Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain
29. Universitat Autònoma de Barcelona, Spain
30. Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA
31. UW Medicine, Seattle, WA, USA
32. Observational Health Data Analytics, Janssen Research and Development, Raritan, NJ, USA
33. Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY 10032, USA
34. New York-Presbyterian Hospital, 622 W 168 St, PH20 New York, NY 10032 USA
35. Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, USA
36. VINCI, Tennessee Valley Healthcare System VA, Nashville, TN & Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN
37. Tufts Institute for Clinical Research and Health Policy Studies, US
38. Real World Solutions, IQVIA, Cambridge, MA, USA
39. Centre for Statistics in Medicine (CSM), Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDROMS), University of Oxford, UK

Corresponding Author
Asieh Golozar
Regeneron Pharmaceutical, NY USA
Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, MD USA
Tel: +1(410)302-6641
email: asieh.golzoar@gmail.com

ABSTRACT

Early identification of symptoms and comorbidities most predictive of COVID-19 is critical to identify infection, guide policies to effectively contain the pandemic, and improve health systems' response. Here, we characterised socio-demographics and comorbidity in 3,316,107persons tested and 219,072 persons tested positive for SARS-CoV-2 since January 2020, and their key health outcomes in the month following the first positive test. Routine care data from primary care electronic health records (EHR) from Spain, hospital EHR from the United States (US), and claims data from South Korea and the US were used. The majority of study participants were women aged 18-65 years old. Positive/tested ratio varied greatly geographically (2.2:100 to 31.2:100) and over time (from 50:100 in February-April to 6.8:100 in May-June). Fever, cough and dyspnoea were the most common symptoms at presentation. Between 4%-38% required admission and 1-10.5% died within a month from their first positive test. Observed disparity in testing practices led to variable baseline characteristics and outcomes, both nationally (US) and internationally. Our findings highlight the importance of large scale characterization of COVID-19 international cohorts to inform planning and resource allocation including testing as countries face a second wave.

INTRODUCTION

As of 22nd October 2020, more than 41 million confirmed cases of coronavirus disease 2019 (COVID-19), and more than 1.1 million deaths have been reported worldwide[1]. As many countries emerge from the impact of the initial phase of the COVID-19 pandemic, strategies to contain the virus are rooted in the rapid testing and identification ('test and trace'), and subsequent isolation (quarantine) of emerging cases to mitigate community transmission and future outbreaks[2].

Great heterogeneity exists in testing policies worldwide, with some countries performing population-scale mass testing, and others opting for more selective testing approaches. Testing availability has also changed rapidly over time. Initially, testing in many countries was limited to only severe cases or those with symptoms, but that has rapidly changed as resources and testing capabilities have become more widely available[3]. Spain, for example, had most of the testing carried out in hospitals in the initial phase of the pandemic, but has now implemented primary care and community mass screening. However, it is likely that testing capacity will be insufficient again as second waves of COVID-19 arise in different countries. This is becoming obvious in the United Kingdom (UK) and the United States (US) amongst others, where tests are sometimes unavailable and/or delayed beyond what is needed for effective test-tracing[4, 5].

Although aggregated figures are updated regularly and in the public domain, there is a scarcity of linked patient-level data on the clinical characteristics and outcomes of those tested and those tested positive[1, 6]. While many previous reports have described people tested positive or with a confirmed diagnosis of COVID-19, only a minority focused on population-based cohorts. In addition, most of such literature consists of case reports or case series, with only a very small minority including at least 1,000 participants.

Understanding the characteristics at the time of testing and health outcomes following a positive test is crucial, not only to understand disease severity and spectrum of illness, but also to provide information for forecasting COVID-19 spread and healthcare provision in the coming months. In this study, we aimed to characterize baseline socio-demographics, comorbidities and symptoms seen at the time of testing, and to assess key health outcomes amongst those tested and those tested positive.

METHODS

Study Design

We performed a large-scale cohort study across a network of outpatient and inpatient electronic health records (EHRs) and national claims data, all standardized to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)[7]. The OMOP CDM and associated tools were used to conduct large scale federated analytics within the Observational Health Data Sciences and Informatics (OHDSI) global research network, whereby each participating institution retained their patient-level data, and run analytical code made available to them by analysts for remote querying, i.e. in a distributed network fashion[8]. This study is part of the "Characterizing Health Associated Risks, and Your Baseline Disease In SARS-COV-2 (CHARYBDIS)" study. The study protocol is available at https://github.com/ohdsi-studies/Covid19CharacterizationCharybdis/blob/master/documents/Protocol_COVID-19%20Charybdis%20Characterisation_V5.docx

Data

Of twenty-one databases contributing data to OHDSI's open science community efforts to characterize the natural history of COVID-19, 12 were included in this study. Nine data sources were excluded due to absence of testing data, insufficient (<1 year) follow-up before testing or where test data were unavailable. Included data in the study were obtained from US health systems including

Columbia University Irving Medical Center (CUIMC) in New York (NY), STAnford medicine Research data Repository (STARR-OMOP) in California (CA)[9], Tufts CLARET (TRDW) in Massachusetts (MA), University of Colorado Anschuz Medical Campus Health Data Compass (CU-AMC HDC) in Colorado (CO) and UW Medicine COVID Research Dataset (UWM-CRD) in Washington (WA) state; US-wide hospital data from Premier, Optum® de-identified COVID-19 Electronic Health Record dataset (Optum EHR), and the Department of Veterans Affairs (VA-OMOP)[10]; and national claims data from IQVIA Open Claims and HealthVerity. Data from South Korea (SK) came from nation-wide claims recorded in the Health Insurance Review & Assessment (HIRA) database. Spanish data came from the Information System for Research in Primary Care (SIDIAP), a primary care database from Catalonia linked to hospital admissions and regional COVID-19 tests. Data collection period varies by database from January 2020 to the end of data collection. Most data were collected from March to May 2020, whilst the data collection period for five databases span to June 2020 and beyond (CUIMC, HealthVerity, STARR-OMOP, UWM-CRD and VA-OMOP). Details on the included data sources are available in Supplementary Figure 1 and Supplementary Tables 1 and 2.

Study Participants

Two cohorts were studied: 1) individuals with a first real-time PCR SARS-CoV-2 test performed between January 2020 and June 2020 (*tested* cohort), 2) individuals who tested positive for SARS-CoV-2 for the first time during the same timeframe (*tested+* cohort). The respective index dates were the day of the first test for the *tested* cohort, and the test sample collection date of the first positive test for the *tested+* cohort. Cohort participants were followed from index date to the earliest of death, end of the study period, or 30 days after index. The codes used to identify COVID-19 diagnoses and SARS-CoV-2 test are described in more detail in Supplementary Table 3. All study participants were required to have at least 365 days of prior observation time before index date to allow comprehensive capture of comorbidities and medication use.

Having a negative test for SARS-CoV-2 is a transient status and dependent on the timeframe of the study. A person tested negative for SARS-CoV-2 could test positive for SARS-CoV-2 later in time and be considered a part of the tested positive cohort if the new test was performed during the timeframe of the study. This particularly makes it difficult to study this cohort during the pandemic. As such, we did not report on the symptoms and characteristics of persons tested negative for SARS-CoV-2 test.

Baseline characteristics and outcomes of interest

Demographic data including age, sex and symptoms at index date were obtained. Comorbidities from day -30 to day -1 before index date were identified based on the SNOMED CT hierarchy, with all descendant codes included. The key conditions reported here are based on recent reports of associations with COVID-19 infection or outcomes[11]. All conditions are reported in full at an interactive website, although this is dynamic and will change as new participants and/or databases are added over time. Detailed definitions of each condition are available at http://atlas-covid19.ohdsi.org/#/home.

The main study outcomes were ascertained from index to day 30 post-index and included hospitalization and death. Additional respiratory, cardiovascular (i.e. acute myocardial infarction, sudden cardiac death, ischemic stroke, intracranial bleed and heart failure), Venous thromboembolism events (i.e. pulmonary embolism and deep vein thrombosis) and other secondary outcomes were also analyzed.

Statistical analyses

All features and outcomes were summarized as proportions and reported per database separately. Databases were grouped into primary care EHR, hospital (inpatient and outpatient) EHR, and claims for reporting and plotting purposes to highlight the differences in the nature of the source of data.

All analytical code was developed centrally and run locally in each database in a federated manner, and is available at https://github.com/ohdsi-studies/Covid19CharacterizationCharybdis. The data reported here were extracted from the CHARYBDIS results set (available here in full) on 04/10/2020.

All the data partners obtained Institutional Review Board (IRB) approval or exemption, as appropriate, to conduct this study. Co-authors with access to the relevant documentation and with direct access to each of the contributing data sources are mentioned in Supplementary Table 1.

RESULTS

A total of 3,316,107 persons were tested and 219,072 tested positive in the participating databases between January and June 2020 and were therefore included in the *tested* and *tested+* cohorts respectively. Twelve databases (one primary care, eight hospital EHR, and three health claims databases) from three countries (Spain, SK and the US) were included (Supplementary Figure 1). For databases with both testing and test result data available (eight out of twelve), the ratio of positive/tested varied dramatically, from 2.3:100 in STARR-OMOP, 6.2 in VA-OMOP, 13.4:100 in HealthVerity, 22.5:100 in SIDIAP, and up to 31.2:100 in CUIMC. This ratio decreased over time, with much higher positive/tested ratios in February-April (50:100 in April) than in May-June (6.8:100 in June).

Baseline demographics, comorbidities, and symptoms:

The majority of study participants were aged 18-64 years old, with small representation of children and a variable proportion of elderly (>65) ranging from 20% to 48% of the total *tested*. Women were overrepresented amongst the *tested* (52% in HIRA to 64% in SIDIAP) in all data sources except VA-OMOP, which is a predominantly male cohort. Detailed patient characteristics including socio-demographics, pre-existing comorbidity and prior drug use are reported in Table 1.

Key comorbidities were generally more prevalent in the *tested+* compared to the overall *tested* (Figure 1, and age- and sex-stratified in Supplementary Figures 2). These differences were consistent across databases for hypertension (59.9% in the *tested+* vs 20.1% in the *tested* in VA-OMOP), obesity (44.4% vs 31.2% in VA_OMOP) and heart disease (42.4% vs 18.8% in VA-OMOP).

The prevalence of symptoms at index date in those *tested* and *tested+* is depicted in Figure 2 (stratified by age and sex in Supplementary Figures 3). The most commonly reported symptoms amongst the *tested* were cough (2.6% in UWM-CRD to 27.5% in IQVIA-OpenClaims), fever (from 1.1% in SIDIAP to 18.6% in HIRA), and dyspnoea (1.3% in SIDIAP to 15.1% in TRDW). The prevalence of these symptoms was higher in those *tested+* and in hospital EHR compared to claims or primary care databases.

30-day outcomes amongst the tested+:

The percentage of patients hospitalized within 30 days after testing positive was: 4.2% (HealthVerity), 9.6% (STARR-OMOP), 19.5% (UWM-CRD), 19.8% (Optum-EHR), 22.9% (TRDW), 25.5% (VA-OMOP), 27.1% (SIDIAP), and 37.6% (CUIMC). Overall fatality in the month after the first positive test ranged from 9.2% (CUIMC) to 10.5% (SIDIAP) (Figure 3). There was a general downward trend over time in outcomes. For example, from March to May, hospitalization decreased from 45% to 14.3% and 30-day mortality decreased from 11% to 1.2% in CUIMC.

Regarding other complications, the percentage of *tested+* patients diagnosed with pneumonia within 30 days from the test date ranged from 3.6% (HealthVerity) to 22.4% (VA-OMOP); acute respiratory distress syndrome (ARDS) from 1.1% (HealthVerity) to 12.0% (VA-OMOP); sepsis 0.6% (HealthVerity) to 4.9% (VA-OMOP). Acute kidney injury was the most common renal outcome (0.6% in HealthVerity to 7.9% in VA-OMOP), followed by dialysis (0.2% in HealthVerity to 1.5% in both TRDW and VA-OMOP). The composite of total cardiovascular events occurred in 0.2% (SIDIAP) to 5.0% (VA-OMOP),

whilst venous thromboembolic events were recorded in 0.2% (HealthVerity) to 2.3% (VA-OMOP) (Supplementary Figure 4).

DISCUSSION

Including over 3.32 million tested and over 219,000 tested positive, we report on the largest patient-level cohort to date of individuals tested for SARS-CoV-2 with linked baseline characteristics and health outcomes. We conducted our analyses in 12 real world data sources from three continents. Positive test ratios varied widely across geography from 2:100 in California to >25:100 in Spain and >30:100 in NYC, illustrating great heterogeneity in testing practice and extent of testing. Similar variability was observed over time in the first half of 2020, with much higher positive/test ratios in February-March, and lower in May-June.

Notwithstanding the known effect of age and male sex on COVID-19 severity, we show that a majority of those tested and *tested+* were in fact adult (18-65 years) women. This is consistent with official sources from the US[12], Spain[13], and SK[14]. Plausible explanations include higher work-related exposure (i.e. hospital staff, care homes)[15-17] and/or differences in healthcare seeking behaviour leading to increased testing in women[18].

The most common comorbidities reported in previous systematic reviews were hypertension, cardiovascular disease and diabetes[19, 20]. In line with previous knowledge, hypertension, obesity and cardiovascular disease were the most prevalent comorbidities in the *tested+* cohort with over 34% of the population suffering from at least one of these three conditions in the US[21], which was consistent with our findings. The higher proportion of these chronic comorbid conditions can be due to the targeted testing of affected people, or a potential association between these conditions (or related treatments) and an increased susceptibility to SARS-CoV-2 infection[22, 23].

The identified range of symptoms is wide, and coding differs across settings, with more symptoms being recorded in hospital data sources than primary care data such as SIDIAP. Despite this variability, there is a consistent recording of fever, cough and dyspnoea as the most common presenting symptoms at the time of testing. All of these symptoms are more common amongst the *tested+* than in the overall *tested* population. Despite under-reporting of symptoms in our study, our findings are consistent with other published literature, including reports from the Centers for Disease Control and Prevention (CDC)[24] and the World Health Organization (WHO)[25]. This lower prevalence compared to research cohorts may be due to incomplete capture of symptoms in coded/structured records. More work is needed to further unravel the potential discriminatory ability of these symptoms, alone or combined, for the differential diagnosis of COVID-19 in the presence of competing influenza and other viruses causing influenza-like disease.

Finally, we studied a number of health outcomes associated with COVID-19 infection. Overall, 4% to 38% of the *tested+* cohort required hospitalization, and 4% to 11% died within a month of their first positive test. Other commonly observed outcomes of interest included pneumonia, ARDS, cardiovascular events, acute kidney injury and venous thromboembolic events. One report from the CDC published at the end of March 2020 reported that in the US, 21-31% of those *tested* were hospitalized[26]. In Europe, a report from the European Center for Disease Prevention and Control (ECDC) published in April 2020 reported that among all cases, 32% were hospitalized[27]. These proportions are likely to decline over time as less severe cases are increasingly tested with better testing capabilities. The fatality proportions reported in our study were slightly higher than the latest published data from the John Hopkins Coronavirus Resource Center (at 3% for both the US and Spain)[28], largely because most of our data were from March to May 2020, with the outbreak at its peak. We also compared our results with the CDC report published at the end of May[29], in which the reported percentage of deaths attributed to pneumonia, influenza or COVID-19 was 13.7%.

Strengths and limitations

Our study has limitations. We analysed data collected during routine care and in actual practice settings for clinical rather than research purposes. It is therefore expected that some information will be incompletely reported, leading to potential misclassification. The variability observed in testing practices and the similarity in the distribution of important comorbidities, symptoms and demographics in the *tested* vs *tested+* cohorts suggests targeted testing for subjects with severe forms of disease in the first months of the pandemic. The lack of mild infections limits our ability to identify key variables associated with susceptibility to infection. Underreporting of symptoms in EHR is another limitation of the study which can be due to factors such as source of data, testing availability, model of care, and reimbursement policies. We observed wide-spread variation in symptom reporting and a generally lower prevalence of symptoms compared to previous literature. Despite this, our study suggests that symptoms such as cough and fever remain key disease features, predictive of a positive test. We were not able to describe characteristics of the population tested negative for SARS-CoV-2 given the transient state of this population and specifically the difficulties in accurately ascertaining this cohort during a pandemic. However, the majority of the *tested* population consists of persons who tested negative for SARS-CoV-2 with *tested+* cohort comprising at most 6.7% of the *tested* cohort. As such, the observed difference in characteristics and outcomes between the *tested* and *tested+* cohorts are largely explained by the tested negative population.

This study also has strengths. First, this is the largest patient-level cohort of COVID-19 tested and *tested+* patients to date. This allowed us to study the observed risks of relatively uncommon outcomes, otherwise not identifiable in smaller datasets. Secondly, we obtained data from multiple centres, three countries, three continents, and covering almost 6 months of 2020, allowing for a comprehensive characterisation of the study population, their key baseline characteristics, symptoms and 30-day outcomes. Finally, the use of routinely collected data from multiple sources maximized the external validity and generalizability of our findings to our settings internationally, whilst minimising Hawthorne effects and selection bias.

Conclusions

In the largest cohort study to date, adult women were the majority of the almost 3.32 million people tested, and of the >219,000 tested positive study participants. Fever, cough and dyspnoea were the most common symptoms, more frequently reported in patients tested positive for SARS-CoV-2. The observed heterogeneity in testing practices complicated an accurate measurement of health outcomes related to COVID-19 in the first half of 2020. With increasing testing capacity, as recommended by the WHO[30], the test+/tested ratio declined in May and June compared to March/April. Ensuring sufficient testing capacity will remain a challenge during the current second wave. If achieved, keeping a low (e.g. <5%) test+/tested ratio will help control and understand the COVID-19 pandemic in the coming months.

REFERENCES

1. John Hopkins University. COVID-19 Dashboard by the Center for Systems Science and Engineering at Johns Hopkins University 2020 [Available from: https://coronavirus.jhu.edu/map.html accessed September 26 2020.
2. World Health Organization (WHO). Critical preparedness, readiness and response actions for COVID-19: Interim guidance. Updated June 24, 2020. 2020 [Available from: https://www.who.int/publications/i/item/critical-preparedness-readiness-and-response-actions-for-covid-19 accessed September 26 2020.
3. Max R, Hannah R, Esteban O, et al. COVID-19 testing policies. Published online at OurWorldInData.org 2020 [Available from: https://ourworldindata.org/grapher/covid-19-testing-policy accessed September 24 2020.

4. Mailonline. What's really behind Britain's testing fiasco? 2020 [Available from: https://www.dailymail.co.uk/news/article-8734923/Whats-REALLY-Britains-coronavirus-testing-fiasco.html accessed September 24 2020.
5. 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 [published online ahead of print, 2020 Jul 28]. *J Am Med Inform Assoc 2020;ocaa188* 2020.
6. Max R, Hannah R, Esteban O, et al. Coronavirus Pandemic (COVID-19). Published online at OurWorldInData.org. 2020 [Available from: https://ourworldindata.org/coronavirus accessed September 24 2020.
7. 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):574-78.
8. Platt RW, Platt R, Brown JS, et al. How pharmacoepidemiology networks can manage distributed analyses to improve replicability and transparency and minimize bias [published online ahead of print, 2019 Jan 15]. *Pharmacoepidemiol Drug Saf* 2019:10.1002/pds.4722. doi: 10.1002/pds.4722
9. Datta S, Posada J, Olson G, et al. A new paradigm for accelerating clinical data science at Stanford Medicine. *arXiv* 2020
10. Lynch KE, Deppen SA, DuVall SL, et al. Incrementally Transforming Electronic Medical Records into the Observational Medical Outcomes Partnership Common Data Model: A Multidimensional Quality Assurance Approach. *Applied clinical informatics* 2019;10(5):794-803. doi: 10.1055/s-0039-1697598 [published Online First: 2019/10/24]
11. Burn E, You SC, Sena AG, et al. Deep phenotyping of 34,128 patients hospitalised with COVID-19 and a comparison with 81,596 influenza patients in America, Europe and Asia: an international network study. *medRxiv 2020042220074336* 2020. (accessed September 24, 2020).
12. Centers for Disease Control and Prevention (CDC). CDC COVID data Tracker. Demographic Trends of COVID-19 cases and deaths in the US reported to CDC. 2020 [Available from: https://covid.cdc.gov/covid-data-tracker/?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F2019-ncov%2Fcases-updates%2Fcases-in-us.html#demographics accessed September 24 2020.
13. National Epidemiological Centre (Spain). Report No. 37. Situation of COVID-19 in Spain. COVID-19 report. Updated on July 30, 2020 2020 [Available from: https://www.isciii.es/QueHacemos/Servicios/VigilanciaSaludPublicaRENAVE/EnfermedadesTransmisibles/Documents/INFORMES/Informes%20COVID-19/Informe%20n%C2%BA%2037.Situaci%C3%B3n%20de%20COVID-19%20en%20Espa%C3%B1a%20a%2030%20de%20julio%20de%202020.pdf accessed September 24 2020.
14. Korea Disease Control and Prevention Agency (KDCA). Coronavirus Infectious Disease-19 Outbreak in Korea: Domestic Occurrence Status 2020 [Available from: http://ncov.mohw.go.kr/bdBoardList_Real.do?brdId=1&brdGubun=11&ncvContSeq=&contSeq=&board_id=&gubun= accessed September 24 2020.
15. Boniol M, McIsaac M, Xu L, et al. Gender equity in the health workforce: analysis of 104 countries. Geneva: World Health Organization 2020 [Available from: https://apps.who.int/iris/handle/10665/311314 accessed September 24 2020.
16. World Health Organization (WHO). Delivered by women, led by men: a gender and equity analysis of the global health and social workforce. 2019 [accessed September 24 2020.
17. World Health Organization (WHO). Gender and COVID-19 2020 [Available from: https://apps.who.int/iris/bitstream/handle/10665/332080/WHO-2019-nCoV-Advocacy_brief-Gender-2020.1-eng.pdf?sequence=1&isAllowed=y accessed September 24 2020.

18. Alsan M, Stantcheva S, Yang D, et al. Disparities in Coronavirus 2019 Reported Incidence, Knowledge, and Behavior Among US Adults. *JAMA Netw Open* 2020; 3(6). http://europepmc.org/abstract/MED/32556260 (accessed September 24, 2020).
19. Yang J, Zheng Y, Gou X, et al. Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. *Int J Infect Dis* 2020; 94.
20. Morgan SG , Daniel Sehayek , Sofianne Gabrielli, et al. COVID-19 and comorbidities: a systematic review and meta-analysis. *Postgraduate Medicine* 2020. (accessed September 24, 2020).
21. Centers for Disease Control and Prevention (CDC). COVIDView: a weekly surveillance summary of US COVID-19 activity. Key updates for week 40, ending October 3rd, 2020 2020 [Available from: https://gis.cdc.gov/grasp/COVIDNet/COVID19_5.html accessed October 12 2020.
22. Lockhart SM, O'Rahilly S. When Two Pandemics Meet: Why Is Obesity Associated with Increased COVID-19 Mortality? *Med (N Y)* 2020:10.1016/j.medj.2020.06.005. doi: 10.1016/j.medj.2020.06.005
23. Verdecchia P, Cavallini C, Spanevello A, et al. COVID-19: ACE2centric Infective Disease? *Hypertension (Dallas, Tex : 1979)* 2020;76(2):294-99. doi: 10.1161/hypertensionaha.120.15353 [published Online First: 2020/06/02]
24. Burke RM, Killerby ME, Newton S, et al. Symptom Profiles of a Convenience Sample of Patients with COVID-19 — United States, January–April 2020. 2020; 69. (accessed September 24, 2020).
25. World Health Organization (WHO). Report of the WHO-China Joint Mission on Coronavirus Disease 2019 (COVID-19) 2020 [Available from: https://www.who.int/docs/default-source/coronaviruse/who-china-joint-mission-on-covid-19-final-report.pdf accessed September 24 2020.
26. Centers for Disease Control and Prevention (CDC). Severe Outcomes Among Patients with Coronavirus Disease 2019 (COVID-19) — United States, February 12–March 16, 2020. *MMWR Morb Mortal Wkly Rep* 2020; (69).
27. European Centre for Disease Prevention and Control (ECDC). Coronavirus disease 2019 (COVID-19) in the EU/EEA and the UK - eighth update. 2020 [Available from: https://www.ecdc.europa.eu/sites/default/files/documents/covid-19-rapid-risk-assessment-coronavirus-disease-2019-eighth-update-8-april-2020.pdf accessed September 24 2020.
28. John Hopkins Coronavirus Resource Center. Mortality Analyses 2020 [Available from: https://coronavirus.jhu.edu/data/mortality accessed October 11 2020.
29. Centers for Disease Control and Prevention (CDC). COVIDView: a weekly surveillance summary of US COVID-19 activity. COVIDView summary ending on May 30, 2020 2020 [Available from: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/covidview/past-reports/06052020.html accessed September 24 2020.
30. World Health Organization (WHO). Laboratory testing strategy recommendations for COVID-19: Interim guidance 21 March 2020 2020 [Available from: https://apps.who.int/iris/rest/bitstreams/1272560/retrieve accessed October 2020.

Table 1: Characteristics of people tested with real-time PCR and patients tested positive for SARS-CoV-2 in a network of databases from Spain, South Korea and United States

		Primary care linked EHR		Hospital based EHR													Nationwide Claims			
		SIDIAP		CUIMC		CDM Premier	CU-AMC HDC	Optum-EHR		STARR		TRDW*		UWM-CRD		VA-OMOP		HealthVerity		IQVIA OpenClaims	HIRA
		first test	first + test	first test	first + test	first test	first test	first test	first + test	first test	first + test	first test	first + test	first test	first + test	first test	first + test	first test	first + test	first test	first test
Total, N		171,810	38,657	18,053	5,625	2,304	99,112	835,797	57,381	39,877	902	3,719	520	53,581	1,777	501,106	31,293	620,962	82,917	739,518	230,268
Index Date*, %	**Jan**	NA	NA	NA	NA	0.00	NA	0.00	NA	NA	NA			NA	NA	NA	NA	NA	NA	NA	0.00
	Feb	0.03	NA	NA	NA	1.65	NA	0.00	NA	2.90	8.43			NA	NA	0.00	NA	NA	NA	0.00	23.74
	March	25.99	47.43	17.50	27.93	20.23	4.34	4.22	12.25	13.69	30.82			8.81	23.41	3.16	7.54	18.20	22.13	27.80	58.73
	April	64.14	50.97	41.69	53.14	38.63	10.37	10.91	20.68	30.48	31.93			12.88	26.22	14.37	14.46	44.34	52.88	54.56	17.03
	May	9.84	1.59	40.81	18.93	36.02	22.88	17.10	13.84	35.29	17.74			18.66	8.55	14.16	7.98	26.98	19.61	17.63	0.37
	June	NA	NA	NA	NA	3.43	29.30	20.39	11.53	16.32	10.42			18.96	8.27	20.96	17.16	10.47	5.38	NA	NA
Age, %	**< 18**	1.68	0.25	2.85	1.33	2.65	3.22	6.67	4.88	6.07	4.88	3.71	1.92	1.67	2.64	0.00	0.00	3.17	2.23	5.39	6.15
	18-64	63.24	51.37	70.60	60.78	49.22	72.04	69.96	76.27	69.07	72.84	75.64	75.96	73.18	72.88	55.37	62.72	77.31	80.26	74.34	70.09
	>= 65	35.08	48.38	26.56	37.88	48.13	24.74	23.37	18.85	24.86	22.28	20.65	22.12	25.15	24.48	44.62	37.29	19.52	17.52	20.27	23.75
Sex, % Female		63.52	59.50	62.48	55.25	55.03	59.63	59.20	56.67	63.11	55.43	55.53	50.38	53.30	50.03	18.85	17.96	59.85	55.95	60.32	52.25
Pre-existing condition of COVID risk factor, %**		34.83	43.16	46.25	52.30	81.94	38.58	45.52	40.22	36.95	30.38	40.44	41.35	37.30	36.80	64.74	59.70	22.62	21.15	46.12	42.45

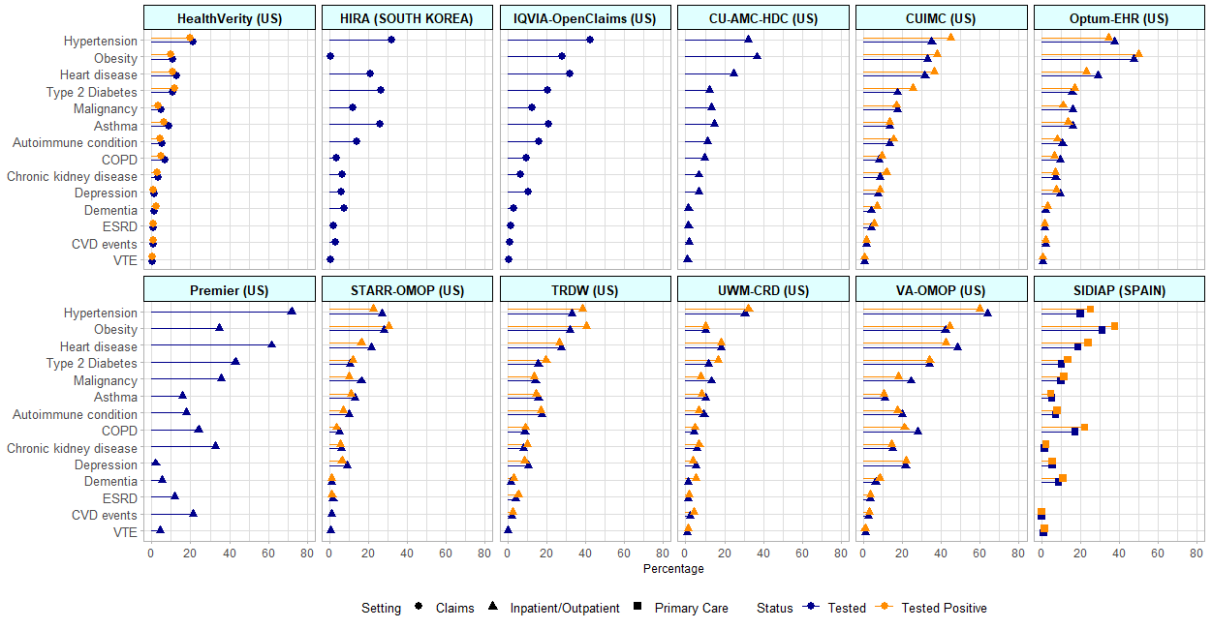
NA= testing was not performed in the respective month

**breakdown of the total counts by index date not available for TRDW. Only data from January to June are presented here.

**any of Chronic kidney disease, stroke (ischemic or haemorrhagic), HIV, malignant neoplasms excluding non-melanoma skin cancer, type 2 Diabetes Mellitus, heat failure, cardiac arrhythmia composite, heart disease condition

Abbreviations: U of Colorado Anschuz Medical Campus Health Data Compass (CU-AMC HDC), Columbia University Irving Medical Center (CUIMC), Health Insurance Review & Assessment Service (HIRA), Optum® de-identified COVID-19 Electronic Health Record dataset (Optum-EHR), Information System for Research in Primary Care (SIDIAP), Tufts Research Data Warehouse (TRDW), UW Medicine COVID Research Dataset (UWM-CRP), Department of Veterans Affairs (VA-OMOP)

Figure 1: Baseline comorbidities 30-days prior to index date among SARS-CoV-2 *tested* and *tested+* cohorts across databases of various setting



COPD = Chronic obstructive pulmonary disease; ESRD = End stage renal disease; CVD = Cardiovascular disease; VTE = Venous thromboembolism events; US= United States

Figure 2: COVID-19 symptoms at index date among SARS-CoV-2 *tested* and *tested+* cohorts across databases of various setting

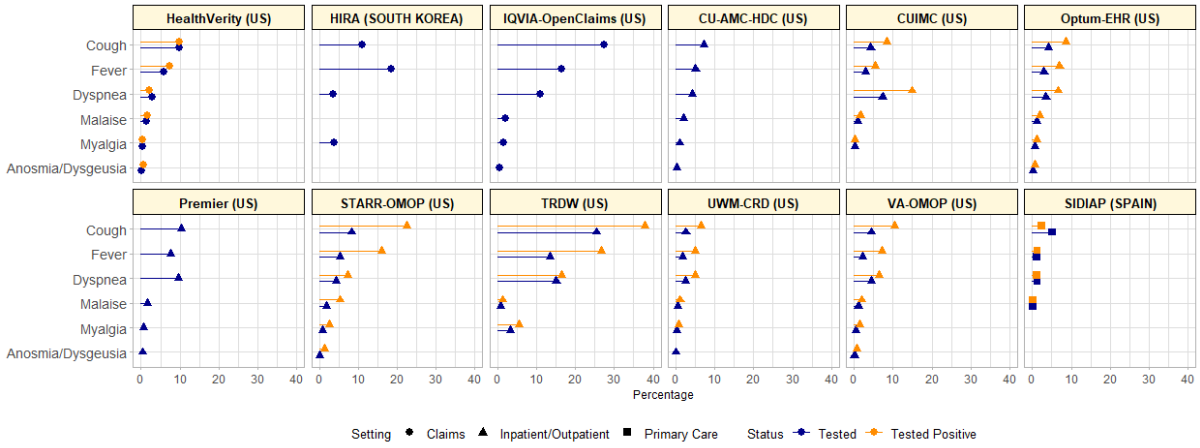
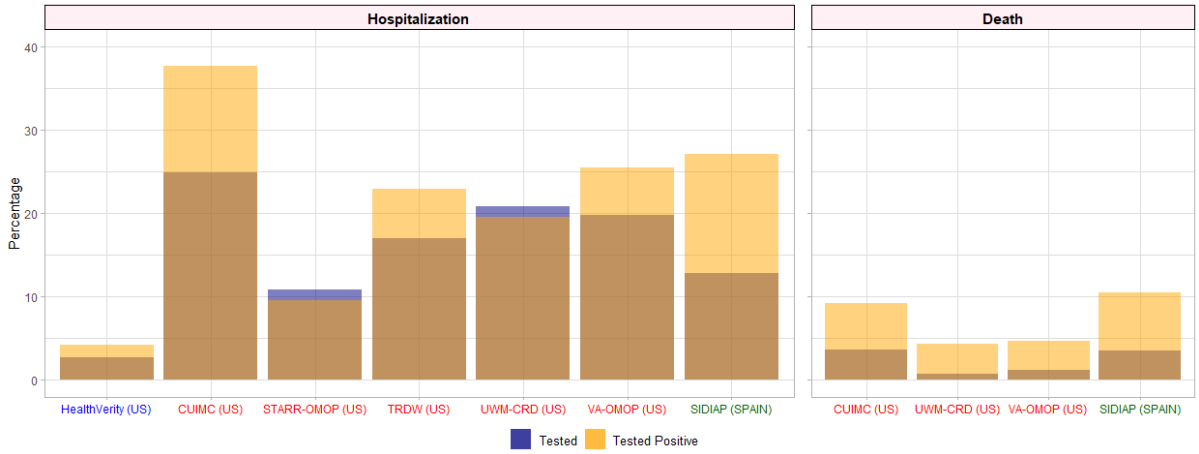


Figure 3: Hospitalization and death 30 days after first test or first positive test, respectively, among SARS-CoV-2 *tested* and *tested+* cohorts across databases of various setting



The overlapped area (brown) indicates an overlap between both tested and tested+ cohorts. Different coloured fonts indicate settings. Blue = Claims; Red =Hospital EHR; Dark green = Primary care EHR

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, 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). DPA received funding from NIHR in the form of a Senior Research Fellowship (SRF-2018-11-ST2-004). WURA reports funding from the NIHR Oxford Biomedical Research Centre (BRC), Aziz Foundation, Wolfson Foundation, and the Royal College Surgeons of England. 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, or the Department of Health, England. UW Medicine received a grant related to this work from the Bill & Melinda Gates Foundation (INV-016910).

ETHICAL CONSIDERATIONS

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-OMOP 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 CPRD was approved by the Independent Scientific Advisory Committee (ISAC) (protocol number 20_059RA2). The use of IQVIA-OpenClaims was exempted from IRB approval. The CU-AMC study participation was considered exempt by the Colorado Multiple Institutional Review Board. The use of the UWM-CRD data was reviewed and approved by the University of Washington Institutional Review Board (STUDY 00010708).

CONTRIBUTORSHIP STATEMENT

TDS, KK, APU, PR, DPA, FN, TMA, LMS, and AG conceived and designed the study. SLD, TF, KEL, MEM, KN, JDP, CGR, NHS, PR, KK, LMS, JAT, ABW and TDS coordinated data contributions at their respective sites. AP, AGS, TF, SFB, JDP, KK and TDS analyzed the data; AG and LL produced the figures and tables. AG, LL, DPA, JJ, FN, TMA. interpreted the data. AG, LL, DPA searched the literature and wrote the first draft. LL, AG, DV, CA, WA, PC, OA, JJ, VS, and HA reviewed the literature and revised the manuscript. All authors contributed to the revision of the first draft, reviewed and approved the final version of the manuscript.

COMPETING INTEREST STATEMENT

All authors have completed the ICMJE uniform disclosure form, with the following declarations made: DPA reports grants from Amgen, grants and other from UCB Biopharma, grants from Johnson and Johnson, outside the submitted work. DM 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), and Tenovus outside the submitted work. SCY reports grants from Korean Ministry of Health & Welfare, grants from Korean Ministry of Trade, Industry & Energy, during the conduct of the study. AG reports personal fees from Regeneron Pharmaceuticals, outside the submitted work; and she is a full-time employee at Regeneron Pharmaceuticals. This work was not conducted at Regeneron Pharmaceuticals. Dr. Lane reports grants from Versus Arthritis, grants from Medical Research Council, outside the submitted work. MS reports grants from US National Science Foundation, grants from US National Institutes of Health, grants from IQVIA, personal fees from Janssen Research and Development, during the conduct of the study. HA reports personal fees from Eli Lilly and Company, outside the submitted work. AS reports personal fees from Janssen Research & Development, during the conduct of the study; personal fees from Janssen Research & Development, outside the submitted work. AS is a full-time employee of Janssen and shareholder of Johnson & Johnson. FD reports personal fees from Janssen Research & Development, during the conduct of the study; personal fees from Janssen Research & Development, outside the submitted work. KK reports she is an employee of IQVIA. CR reports he is an employee of IQVIA. FN was an employee of AstraZeneca until 2019 and holds some AstraZeneca shares. SF is an employee of Janssen Research and Development, a subsidiary of Johnson and Johnson. VS reports grant funding from the National Science Foundation, Agency for Healthcare Research and Quality, and the Arizona Board of Regents outside of the submitted work. 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. PR reports and is employee of Janssen Research and Development and shareholder of Johnson & Johnson. No other relationships or activities that could appear to have influenced the submitted work.

ACKNOWLEDGEMENTS

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