Multicentre collaboration for COVID-19 observational studies Report 1

EMA/198302/2020

30th October 2020

Study:	Multicentre collaboration for COVID-19 observational studies (Report 1)
Prepared for:	European Medicines Agency, Amsterdam, the Netherlands
Prepared by:	E-CORE NETWORK

Table of contents

List of figures

List of tables

Abbreviations

Term	Definition
ACE	Angiotensin-converting-enzyme
AE	Adverse event
AP-HP EDS	The Assistance Publique – Hôpitaux de Paris Health Data Warehouse (AP-HP EDS)
CDM	Common Data Model
COVID-19	Coronavirus disease 2019
DA	Disease Analyser
DQD	Data Quality Dashboard
ECMO	Extracorporeal membrane oxygenation
EHDEN	European Health Data and Evidence Network
EMA	European Medicine Agency
ETL	Extraction transformation Load
GP	General practitioner
HM	Hospital de Madrid
IL	Interleukin
IL-6R	Interleukin-6 receptor
IMI	Innovative medicines initiative
IMRD	IQVIA Medical Research Data
IPCI	Integrated Primary Care Information
JSON	JavaScript Object Notation
LPD	Longitudinal Patient Database
TNF	Tumor necrosis factor
OHDSI	Observational Health Data Science and Informatics
OMOP	Observational Medical Outcomes Partnership
RCT	Randomised clinical trial
SAB	Scientific Advisory Board
SIDIAP	Information System for Research in Primary Care (SIDIAP),

Section 1.0 Executive summary

COVID-19 is an emerging and rapidly evolving infectious disease that has reached pandemic status. It also poses a major global challenge to health-care systems, which have been partially or completely disrupted in many countries due to overwhelming demand.

Currently, no specific anti-viral agent exists for COVID-19. Various drugs such as immunosuppressive and immunomodulatory drugs, antimalarial drugs, Angiotensin-converting-enzyme (ACE) inhibitors and steroids are prescribed to COVID-19 patients off-label. So far, the evidence for the efficacy and safety of the above-mentioned potential treatments for COVID-19 is mostly inconclusive.

Although randomised clinical trials (RCTs) are routinely used to investigate efficacy and safety, they are not well suited for long-term safety and they often exclude at-risk patients. Moreover, they are challenging to conduct in the COVID-19 setting, as institutions are either overwhelmed during outbreaks or lack patients. Therefore, observational, database studies are urgently needed to complement RCTs, especially in the area of safety and for populations excluded from clinical trials.

Observational Health Data Science and Informatics (OHDSI) is a multi-stakeholder, interdisciplinary collaborative network designed to combine the value of health data through large-scale analytics. OHDSI has established an international network of researchers and observational health databases. The OHDSI community has a vast standardised library of codes, methods and programming specifications that can be used to speed up COVID-19 research using real world data.

In June 2020, EMA contracted IQVIA with a project to build a framework for the conduct of multicentre cohort studies on the use of medicines in COVID-19 patients. This project will leverage the resources already constructed by the OHDSI community and aims to accelerate the generation of robust real-world evidence about the utilisation, effectiveness and safety of therapies for COVID-19 treatment.

Report No.1 provides a high-level understanding of eight databases from seven European countries, with a good distribution across Europe. We describe these databases in terms of data quality and completeness, and the applicability to test different potential research objectives. All databases are part of the OHDSI community and have applied a common data model (CDM) or are currently in the process of standardizing their data to a CDM.

We have found that six databases that are currently standardized to the CDM have a median number of 3,600 (min-max range:1,400 to 124,000) COVID-19 patients (either with diagnosis code or with a positive test) as per up to June 2020 (data lock points differ between databases). Sociodemographic variables such as age and sex as well as comorbidities and drug treatment are 100% present in all databases. Other variables such as body mass index (BMI), smoking and type of COVID-19 test (type of different antigen tests) are more sparsely collected: 0-42% for BMI, 0-27% for smoking and 0% for type of COVID-19 tests. Data regarding hospitalisation is present in three databases.

As a next step, we will run a proof of concept study on COVID-19 patients to test the network capabilities and report the results alongside with an evaluation of the collaborative framework.

EMA can use this network to accelerate the generation of evidence about the utilisation, effectiveness and safety of therapies for COVID-19 patients. This will help to improve the understanding of the effectiveness of medicines for better treatment and care for patients with COVID-19.

Section 2.0 Introduction

The Coronavirus disease 2019 (COVID-19) is an emerging and rapidly evolving infectious disease that has reached pandemic status. As of 9 September 2020, more than 27 million people worldwide (~4 million people in Europe) were diagnosed with COVID-19 (ECDC, 2020), whilst the number of deaths has reached over 898,000 worldwide and 212,000 in Europe (ECDC, 2020). COVID-19 poses a major global challenge to health-care systems, which have been partially or completely disrupted in many countries due to overwhelming demand, healthcare workers getting sick and resource diversion (WHO, 2020).

Currently, no specific anti-viral agent for COVID-19 exists. However, immunosuppressive and immunomodulatory drugs are being repurposed with various results. For example, two antimalarial drugs (chloroquine and hydroxychloroquine) showed initial promise but failed to show efficacy in further studies (Torjesen, 2020). Reports on angiotensin-converting-enzyme inhibitors (ACE) inhibitors or angiotensin-II receptor blockers are also inconsistent, i.e. better outcomes for treated patients in some studies (Meng et al., 2020; Zhang et al., 2020), whereas no effect in other studies (Li, Wang, Chen, Zhang, & Deng, 2020). Other classes of drugs that show potential include interleukin-6 receptor (IL-6R) antagonists, interleukin (IL)-1 antagonists, tumor necrosis factor (TNF)-alpha inhibitors, and Janus kinase inhibitors (Sarzi-Puttini et al., 2020).

Systemic steroids are a medication class that showed promising results in a recent randomised clinical trial (RCT) in the UK leading to a reduction of death in COVID-19 patients treated with dexamethasone by 35% in ventilated patients and by 20% amongst patients on supplemental oxygen therapy. However, no benefit was observed in mild cases (Group et al., 2020). So far, the evidence for the efficacy of the above-mentioned potential treatments for COVID-19 is starting to accumulate but it is not yet strong enough to support clear treatment recommendations.

Besides efficacy, drug safety is another major aspect of medical therapy that drives medical decision making. This is especially true when the benefit-risk balance is uncertain and in vulnerable patients. Although RCTs are routinely used to investigate efficacy and safety, the long-term safety outcomes and adverse drug reactions with a low incidence are usually not captured especially and certain at-risk categories of patients are often excluded. Moreover, although the speed of running an RCT increased significantly in the context of the pandemic, they are still challenging to conduct as healthcare institutions tend to either be overwhelmed in areas of surge or lack patients where lockdown measures have managed to control the virus. Therefore, RCTs need to be rapidly complemented by observational database studies, especially in the area of safety and for populations excluded from clinical trials, such as patients seen in primary care. Indeed, the majority of COVID-19 patients are treated in the primary care setting.

Observational database studies using real-world data are underway across Europe and worldwide, employing various study designs across a wide range of patient populations (Singh, Majumdar, Singh, & Misra, 2020). In this environment, multinational research networks dedicated to observational studies are essential. They can provide valuable insight through comparison between different institutions from different countries, and therefore validate true clinical findings from artefacts due to different healthcare settings and data capture modalities.

One of these is the Observational Health Data Science and Informatics (OHDSI) initiative, a multi-stakeholder, interdisciplinary open science collaborative. OHDSI has established an international network of researchers and observational health databases (OHDSI, 2020a). In addition, the European Health Data and Evidence Network (EHDEN) consortium, established under the Innovative Medicines Initiative, is extending the OHDSI network in the European Setting including COVID-10 databases.

The OHDSI community recently applied its network to the COVID-19 'study-a-thon'. Several dozens of stakeholders generate meaningful outcomes using real world data effectively (https://www.ohdsi.org/covid-19-updates/), based on a standard and reusable library of codes, methods and programming specifications (OHDSI, 2020b).

Another important stakeholder, the governmental agencies are also investing resources in COVID-19 research, for example, the European Medicines Agency (EMA) has now set up an infrastructure to support the monitoring of the efficacy and safety of COVID-19 treatments as well as those of vaccines. In June 2020, EMA contracted IQVIA with a project to build a framework for the conduct of multicentre cohort studies on the use of medicines in COVID-19 patients.

We brought together a consortium of data partners, academia, epidemiologist, and data scientists that collaboratively aim to accelerate the generation of robust real-world evidence about the utilisation, effectiveness and safety of therapies for COVID-19 patients.

Section 3.0 Aim and objectives

Report No.1 describes the landscape for assessing electronic medical databases from at least seven European countries, to better understand the feasibility of conducting future multicentre observational studies on COVID-19 using electronic healthcare records. This report provides a high-level understanding of each database in terms of data quality and completeness from which the database's value and applicability to different research objectives can be determined.

The specific objectives of this report are to describe:

- The participants of the database network and the availability of patients over time (Section 5)
- The selection criteria for considering a database admissible in the collaboration (Section 6.0)
- The technical solutions to be applied to assess the selected databases for clinical outcomes, relevant lifestyle factors, socio-demographic data, current and past medical history and current and past drug utilisation history (Section 4.0)
- Results on the assessment of already included databases (Section 7.0)

The report is structured following the EMA required objectives as per the technical specification document, Multicentre collaboration for COVID-19 observational studies, EMA/198302/2020.

Section 4.0 Methods

4.1 Enrolling partners in the EMA Covid-19 Network

To identify data sources appropriate to be included in the network, named **EMA Covid-19 Network** we assessed data from two extensive existing networks, OHDSI and EHDEN, which include primary and secondary care settings from seven European countries, see Table 1 below.

Table 1: Characteristics of databases

Database	Managing Organisation	Country	Covered Patient Lives	History
Enrolled and mapped to OMOP CDM				
LPD Belgium	IQVIA	Belgium	1.1M	2005 – present
LPD France	IQVIA	France	7.8M	1994 – present
DA Germany	IQVIA	Germany	34M	1992 – present
UK IMRD	IQVIA	UK	15.2M	1996 – present
LPD Italy	IQVIA	Italy	2M	2004 – present
IPCI	Erasmus MC	Netherlands	2.6M	1996 – present
SIDIAP	IDIAP Jordi Gol	Spain	7.8M	2006 – present
HM Hospitales	IDIAP Jordi Gol	Spain	~17M	2019 – present
Enrolled, mapping to OMOP CDM is ongoing				
Clinical Center Serbia	Clinerion	Serbia	400k	2004 – present
EDS	APHP	France (Paris)	11M	2012 – present

Database	Managing Organisation	Country	Covered Patient Lives	History
Technical University Dresden	Self	Germany	480k	2012 – present
SNDS	Pharmacoepidemiologie CHU-Lyon	France	Unknown	2006-present

Any data asset in a European country can participate in the EMA Covid-19 Network, provided they are compliant with the OMOP CDM. Data partners wishing to contribute have to undergo a process of contracting, CDM transformation and quality control, as described below.

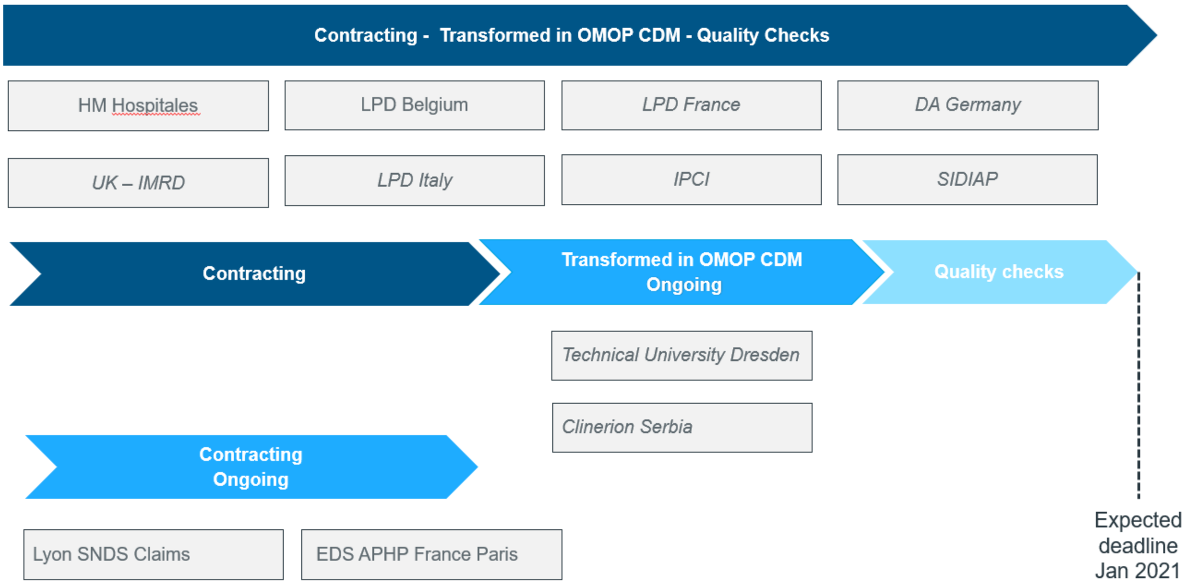


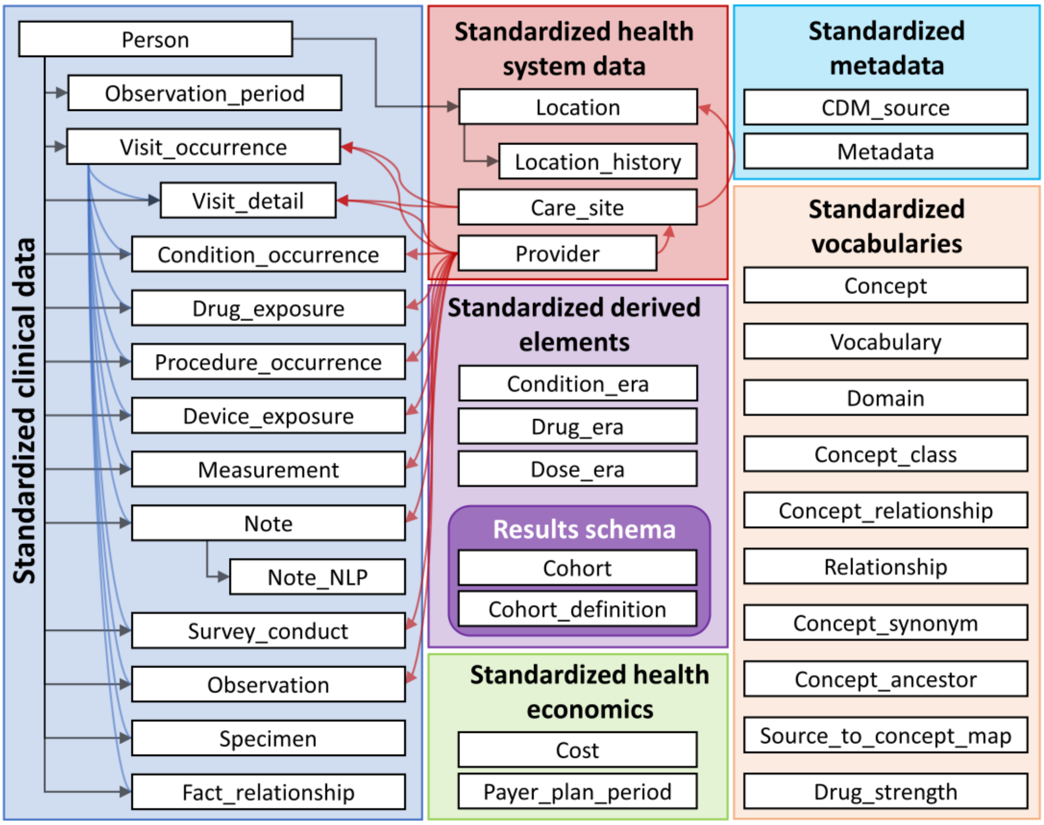
Figure 1: Staged enrolment of data partners over time

4.2 Standardizing databases in the OMOP CDM

To assess and analyse multiple data sources concurrently, data need to be harmonised into a common data standard. In addition, patient data require a high level of protection. A common data standard can alleviate this need by omitting the extraction step and allowing a standardised analytic to be executed on the data in its native environment.

This standard is provided by the CDM. The CDM, combined with its standardised content ensures that research methods can be systematically applied to produce meaningfully comparable and reproducible results.

All databases included in this network will be standardised to the OMOP common data model. The CDM covers the specification for all variables that can be collected throughout the study and enables the use of standardised analytics and tools across the network since the structure of the data and the terminology system is harmonised. The OMOP CDM is developed and maintained by the OHDSI initiative and is described in detail on the wiki page of the CDM: https://ohdsi.github.io/CommonDataModel/ and in The Book of OHDSI: http://book.ohdsi.org.



Each OMOP data partner may execute a Study R package against their database to generate the data. After review of the results the data custodian returns them to the coordinating centre (IQVIA Ltd UK).

An overview of all the tables in the CDM is provided in **Error! Reference source not found.**.

Figure 2: Overview of all tables in the CDM version 5.3

4.3 OMOP CDM Data Quality Checks

OHDSI and EHDEN quality control mechanisms for the Common Data Model were applied. The Data Quality Dashboard (DQD) has been developed in the EHDEN project in close collaboration with OHDSI. and are described in high detail in Chapter 15 of The Book of OHDSI (http://book.ohdsi.org/DataQuality.html) and in the publication (Kahn et al., 2016).

This package runs a series of data quality checks against an OMOP CDM instance (currently supports v5.3.1 and v5.2.2). It systematically runs the checks, evaluates the checks against some pre-specified threshold, and then communicates what was done in a transparent and easily understandable way. The quality checks were organised according to the Kahn Framework. which uses a system of categories and contexts that represent strategies for assessing data quality (Kahn et al., 2016).

Using this framework, the Data Quality Dashboard takes a systematic-based approach to running data checks. Instead of writing thousands of individual checks, we use “data quality check types”. These “check types” are more general, parameterised data quality checks into which OMOP tables, fields, and concepts can be substituted to represent a singular data quality idea.

Version 1 of the tool includes 20 different check types organised into Kahn contexts and categories. Additionally, each data quality check type is considered either a table check, field check, or concept-level check. Table-level checks are those evaluating the table at a high-level without reference to individual fields, or those that span multiple event tables. These include checks making sure required tables are present or that at least some of the people in the PERSON table have records in the event tables. Field-level checks are those related to specific fields in a table. The majority of the check types in version 1 are field-level checks. These include checks evaluating primary key relationship and those investigating if the concepts in a field conform to the specified domain. Concept-level checks are related to individual concepts. These include checks looking for gender-specific concepts in persons of the wrong gender and plausible values for measurement-unit pairs. For a detailed description and definition of each check type, you can refer to the GitHub documentation: https://ohdsi.github.io/DataQualityDashboard/articles/CheckTypeDescriptions.

After systematically applying the 20 check types to an OMOP CDM version approximately 3,351 individual data quality checks are resolved, run against the database, and evaluated based on a pre-specified threshold. The R package then creates a (JavaScript Object Notation) JSON object that is read into an RShiny application to view the results. Results from the DQD may be used to inform supplemental investigations into ETL processes and identify opportunities for enhancement of each local CDM.

4.4 Quantitative database assessment

A quantitative assessment of the databases was performed after their transformation in CDM.

For the already transformed databases (8 out of 12), standardized analytics using R were applied and adapted to the purpose, to conduct the quantitative assessment automatically. The package creates and characterises a selection of cohorts validated as part of the Charybdis project (https://data.ohdsi.org/Covid19CharacterisationCharybdis/) and was shared with the data partners who ran the package. The results were sent to the coordinating centre (IQVIA). If errors were found or additional data needs collection, the package was amended and resent. The results of the databases can be viewed using the following link: https://dqdashboard.iqvia.com/ema_report1/

Our code uses standardised and widely used libraries such as:

- https://github.com/OHDSI/FeatureExtraction
- https://github.com/OHDSI/SqlRender
- https://github.com/OHDSI/CohortDiagnostics

The overall approach taken to conduct the quantitative assessment was as follows:

- COVID-19 cohorts using different diagnosis definitions (tests versus diagnosis codes) were constructed (see section 7.1)

- Assessment of variable coverage: presence of database variables assessed against a list of desired variables for different research objectives (see section 7.2)
- Assessment of data distribution based on a predefined list of desired variables (see Section 7.3)

The second and third points above about the quantitative assessment were assessed in one of the COVID-19 cohorts initially created (the broadest definition was used). In case the COVID-19 cohort was not large enough (<100 patients), we retrieved a cohort of influenza patients to provide the numbers as a surrogate.

An OMOP Scientific Advisory Board was established to provide scientific oversight to the development of the observational studies and lead the coordination of all tasks under this project. This board is comprised of international experts with extensive experience conducting large scale, multicentre observational research in healthcare.

4.5 Qualitative database assessment

A questionnaire regarding any access restrictions to the data and ethics approvals was sent by email. The questionnaire was filled out manually by the data custodians of the included databases.

Section 5.0 Description of the COVID-19 observational studies Network

5.1 Characteristics of databases

From 12 partners who were screened, eight databases were ready to enrol by September 2020.

The eight databases of interest were selected to provide a mix of primary and secondary care setting, although the secondary databases remain underrepresented for now but might be changed by expanding the network. One country contributes both primary care data as well as hospital data. They also provide a good coverage of European countries, giving the researchers the opportunity to explore the question of interest in different healthcare settings and to test the representativeness of their findings. All of them are electronic medical records and a few contain specialist's data in addition to the GP data.

Longitudinal Patient Database (LPD) Belgium (IQVIA)

LPD Belgium is a computerised network of general practitioners (GPs) who contribute to a centralised database of anonymised data of patients with ambulatory visits. Currently, around 300 GPs from 234 practices are contributing to the database covering 1.1M patients from a total of 11.5M Belgians (10.0%). The database covers a time period from 2005 through the present. Observation time is defined by the first and last consultation dates. Drug information is derived from GP prescriptions. Drugs obtained over the counter by the patient outside the prescription system are not reported. No explicit registration or approval is necessary for drug utilisation studies.

Longitudinal Patient Database (LPD) France (IQVIA)

LPD France is a computerised network of physicians including GPs who contribute to a centralised database of anonymised patient EMR. Currently, >1200 GPs from 400 practices are contributing to the database covering 7.8M patients in France. The database covers a time period from 1994 through the present. Observation time is defined by the first and last consultation dates. Drug information is derived from GP prescriptions. Drugs obtained over the counter by the patient outside the prescription system are not reported. No explicit registration or approval is necessary for drug utilisation studies.

Disease Analyser (DA) Germany (IQVIA)

DA Germany is collected from extracts of patient management software used by GPs and specialists practicing in ambulatory care settings. Data coverage includes more than 34M distinct person records out of at total population of 80M (42.5%) in the country and collected from 2,734 providers. Patient visiting more than one provider are not cross identified for data protection reasons and therefore recorded as separate in the system. Dates of service include from 1992 through present. Observation time is defined by the first and last consultation dates. Germany has no mandatory GP system and patient have free choice of specialist. As a result, data are collected from visits to 28.8% General, 13.4% Orthopaedic Surgery, 11.8% Otolaryngology, 11.2% Dermatology, 7.7% Obstetrics/Gynaecology, 6.2% various Neurology and Psychiatry 7.0% Paediatric, 4.6% Urology, 3.7% Cardiology, 3.5% Gastroenterology, 1.5% Pulmonary and 0.7% Rheumatology practices. Drugs are recorded as prescriptions of marketed products. No registration or approval is required for drug utilisation studies.

IQVIA Medical Research Data (IMRD) UK (IQVIA)

IMRD UK is a large database of anonymised electronic medical records collected at Primary Care clinics throughout the UK. Data coverage includes 15.2M patients, 5.6M providers, 793 care sites and more than 5 billion service records, covering 22.5% of a population of 67.5M. Dates of service include from 1996 through present. Quality indicators define the start date for that patient (e.g. each patient's observation period began at the latest of: the patient's registration date, the acceptable mortality recording date of the practice, the Vision date). The end of the observation period is determined by the end date of registration in the database. Drug treatment is recorded as prescriptions. All protocols must be submitted to an independent Scientific Review Committee prior to study conduct.

Longitudinal Patient Database (LPD) Italy (IQVIA)

LPD Italy is comprised of anonymised patient records collected from software used by GPs during an office visit to document patients' clinical records. Data coverage includes over 2M patient records with at least one visit and 119.5M prescription orders across 900 GP practices. Dates of service include from 2004 through present. Observation time is defined by the first and last consultation dates. Drugs are captured as prescription records with product, quantity, dosing directions, strength, indication and date of consultation.

Integrated Primary Care Information (IPCI), The Netherlands

IPCI is collected from EHR records of patients registered with their GPs throughout the Netherlands. The selection of 391 GPs is representative of the entire country. The database contains records from 2.6 million patients out of a Dutch population of 17M (8.2%) starting in 1996. The median follow-up is 2.2 years. The observation period for a patient is determined by the date of registration at the GP and the date of leave/death. All data before the observation period is kept as history data. Drugs are captured as prescription records with product, quantity, dosing directions, strength and indication. Drugs not prescribed in the GP setting might be underreported. Indications are available as diagnoses by the GPs and, indirectly, from secondary care providers but the latter might not be complete. Approval needs to be obtained for each study from the Governance Board. (5) The IPCI database is currently increasing the update frequency because of the COVID-19 pandemic (now updated till March 2020).

Information System for Research in Primary Care (SIDIAP), IDIAP Jordi Gol (Spain)

SIDIAP is collected from EHR records of patients receiving primary care delivered through Primary Care Teams (PCT), consisting of GPs, nurses and non-clinical staff. The Catalan Health Institute manages 286 out of 370 such PCT with a coverage of 5.6M patients, out of 7.8M people in the Catalan population (74%). The database started to collect data in 2006. The mean follow-up is 10 years. The observation period for a patient can be the start of the database (2006), or when a person is assigned to a Catalan Health Institute primary care centre. Date of exit can be when a person is transferred-out to a primary care centre that does not pertain to the Catalan Health Institute, or date of death, or date of end of follow-up in the database. Drug information is available from prescriptions and from dispensing records in pharmacies. Drugs not prescribed in the GP setting might be underreported; and disease diagnoses made at specialist care settings are not included. Studies using SIDIAP data require previous approval by both a Scientific and an Ethics Committee (9).

Hospital de Madrid (HM) Hospitales (Spain)

Hospital de Madrid (HM) Hospitales data are made available through partnership with SIDIAP. The HM Hospitales database covers in-patient care delivered across a network of private hospitals in Spain. HM Hospitales covers more than 17M patients, out of whom a subset will be catalogued and followed for acute care delivered for suspected COVID-19 onset. This database covers more than 2300 confirmed COVID-19 cases and all in-patient hospital care, including the data of admission, conditions, procedures and medicines dispensed in hospital, date of discharge, and date of known death or date of end of follow-up in the database. Studies using HM Hospitales data require review and approval from data custodians at SIDIAP authorised to execute observational network analyses. The number of newly diagnosed patients in each database per month since the outbreak of COVID-19 can be derived from the databases.

Clinerion (Serbia)

Clinerion's Patient Network Explorer (PNEx) network reports present real-time statistics derived from the live deidentified electronic health records (EHRs) of the 30.4 million patients in the global network of partner healthcare organisations (HCOs) in 15 countries worldwide (including Serbia specially to this scope) on the PNEx platform. PNEx allows real-time multi-dimensional query of patient data from the partner HCOs' networked EHR systems, made interoperable by the use of proprietary semantic and ontology methods, from a hybrid cloud- and federated local installation-based platform. Results are available in aggregated and de-aggregated form across HCOs and geographies, without compromising patient data privacy. Clinerion's proprietary technologies comply with international patient privacy and data security regulations.

AP-HP Health Data Warehouse (AP-HP EDS) (France)

The Assistance Publique – Hôpitaux de Paris Health Data Warehouse (AP-HP EDS) collects data from the main clinical information systems of its 39 hospitals based in Paris. The main data collected includes patient demographics, care data, medical documents, prescriptions, physiological, biological and imaging data. In March 2020, the EDS-COVID cohort was born, comprised with data from more than 20,000 COVID positive patients. The inclusion criterion for the EDS-COVID cohort is the performance of a PCR for the coronavirus, the result of which is validated in GLIMS. The data is then gradually enriched with clinical data present within the EDS.

Technical University Hospital Carl Gustav Carus Dresden (Germany)

The University Hospital Dresden with its 20 clinics, four institutes and ten interdisciplinary centres is the city's largest hospital and the only hospital of maximal care in East Saxony. Each year, 67,900 patients receive state-of-the-art medical treatment. With 1,300 in-patient beds and 95 out-patient facilities we offer the whole range of medical service to highest quality standards, treating an average of 60,000 patients per year. Within the hospital there are centres of competence for the treatment of cancer, vascular disease and many more to meet the requirements of an interdisciplinary approach.

SNDS via Pharmacoepidemiologie CHU-Lyon (France)

The French national claims database called "Système National des Données de Santé" (SNDS) is the mains health care claims database in France with individual anonymous information of primary care and secondary care. It includes information from the claims systems in the "Système National d'Information Inter-régime de l'Assurance Maladie" (SNIIRAM). In 2019, SNIIRAM cover currently

98.8% of the French population. The SNIIRAM was started in 2003 with data from the salaried workers scheme. In 2007, hospital data were added from the "Programme de Médicalisation des Systèmes d'Information" (PMSI) and other health insurance scheme. Data are available from 2007.

The SNDS database contains:

- Demographic characteristics: gender, year of birth, month and year of death for those concerned, residence area or region, information on the mutual complementary insurance systems and beneficiary of CMU-C (people socioeconomically disadvantaged are 100% covered).

- Information on healthcare professionals.

- Presence of chronic condition (ALD) with International Classification of Diseases (ICD-10) codes, start and end date of ALD. Patients registered for an ALD benefit from full coverage for all medical expenses related to that condition.

- All non-hospital reimbursed healthcare expenditures with dates and codes (but not the corresponding medical indication nor outcome): recorded and dispensed drugs identified by a unique national registration code (CIP code) and Anatomical Therapeutic chemical classification (ATC code), and number of packs, date of prescription and dispensing, specialty of prescriber, identifier of the pharmacy.

- Some information on date and nature of medical and paramedical interventions, laboratory tests, medical transportations, and number of days of paid sick leaves (Indémnités Journalières).

- Hospital discharge summaries from PMSI: ICD-10 diagnoses codes (main, related, and associated diagnoses) for all hospitalizations, with the date and duration of hospitalization, medical procedures, hospital ward, and cost coding system. Information on in-hospital prescribing only for very expensive drugs not included in hospital diagnosis-related groups.

How does a patient enter and exit the datasource

LPD Belgium, LPD France, DA Germany, LPD Italy a maximum of transaction (record) is used to calculate observation period..Basically, follow-up starts at first healthcare encounter (visit or prescription) and ends at the last encounter.

IMRD UK – there are indicators that define the start date for that patient (e.g. each patient's observation period began at the latest of the patient's registration date, the acceptable mortality recording date of the practice, the Vision date). The end of the observation period is determined by the end date of registration in the database.

IPCI -The observation period for a patient is determined by the date of registration at the GP and the date of leave/death.

SIDIAP-The observation period for a patient can be the start of the database (2006), or when a person is assigned to a Catalan Health Institute primary care centre. Date of exit can be when a person is transferred-out to a primary care centre that does not pertain to the Catalan Health Institute, or date of death, or date of end of follow-up in the database

HM Hospitales – date of admission and data of discharge or death are used for patient entry and exit.

Databases refreshment dates

Data are generally not refreshed continuously over time, but refreshed at certain time intervals. This is due to the fact that OMOP transformations are timely and compute intense, and observational studies analyse cohorts derived from large populations and are therefore robust against small incremental changes of the underlying data. The timing for data updates and lag time between update and access for the current Network members are presented in Table 2.

Table 2: Update frequency and data latency of databases at current Data Partners

Database	Update frequency	Data latency
LPD Belgium	6-monthly	6-8 weeks lag
LPD France	6-monthly	3 weeks lag
DA Germany	6-monthly	6 weeks lag
UK IMRD	6-monthly	6 weeks lag
LPD Italy	3-monthly	6 weeks lag
IPCI	6-monthly	3 months lag
SIDIAP	6-monthly	2-3 months lag
HM Hospitales*	6-monthly	2-3 months lag
Clinical Center Serbia	24 hours	6 weeks lag
EDS France	Unbeknown	6 weeks lag
Technical University Dresden Germany	Unknown	2-3 months lag
SNDS France	Every 2 months	3 months lag

* The partner was not able to confirm further updates after April 2020

Section 6.0 Database selection criteria

The criteria for a data source to be admissible in the COVID-19 collaboration refers to six domains. The criteria are not related to a specific study protocol, and not all databases will be suitable for all protocols, however as a minimum, the data source must contain sufficient COVID-19 patients in either in-patient or ambulatory settings and essential variables needed to conduct COVID -19 related research (e.g., COVID-19 disease history, disease presentation, laboratory test of COVID-19, and treatment of COVID-10 including medicines and highest oxygen therapy)

The criteria for admitting a database in the COVID-19 collaboration are based on a combination of: recommendations from (Hall et al., 2012) , OMOP Data Quality Checks (see Section 6.3) and published literature specific to COVID-19 research.

Data structure

- Population covered: The data source includes a sufficient population in terms of size, coverage and representativeness of country of origin
- Data update: The database is updated sufficient times (at least 6-monthly)
- Ability to link at least outcomes and drug exposure and, optionally laboratory values at patient level.

Longitudinal dimension

- Follow-up time: start and end of follow-up can be identified or inferred
- The average patient follow-up is long enough to allow meaningful research. This might differ from one study to another.
- Continuous and consistent data capture: No major breaks or changes in data collection over time for either individual patients or the whole population during the study observation period.

COVID-19 testing or diagnosis

- The database contains or will contain in the near future (for lagged data) a large enough COVID-19 patients' cohort either in primary or secondary care
- Diagnosis codes or laboratory tests that would allow identification of COVID-19 are captured
- Socio-demographic variables are captured
- Co-morbidities (any) are captured
- Drug treatments (any) are captured

Quality and validation procedures

- On the source data:
  - Data is entered by trained personnel
  - Appropriate general quality checks are routinely completed
- During extraction–transformation-load process
  - During this process several quality checks of the data are performed. The checks are on conformance, completeness and plausibility of values and are performed on relational, temporal and numerical values. For more details please see section OMOP Data Quality Checks

- A Data Quality Dashboard will be created for each database and will be shared publicly
- All issues found have to be solved or considered non-essential before the data can be used.
- Deviations are documented.

Privacy and security

- Compliance with privacy and security policy: All relevant local, regional and national policies been complied with
- No use of identifying information: All direct identifiers are removed or masked.
- No patient-level data are shared outside the site.

Access, Contracting & Ethics

- Willingness to participate and have resources available for transformation into CDM and running studies
- Allows collaboration with third party researchers as EMA
- The access model allows the OMOP conversion and a distributed network access
- An ethics and/or scientific review process in place for the study protocols

Optional

Previous involvement in research and database expertise

- Expertise required to use the resource available is available in house or externally
- Publications citing the use of the database

Section 7.0 Results of databases assessment

7.1 COVID-19 patient cohorts

In each of the eight included databases that were enrolled and mapped to OMOP CDM, we defined the following COVID-19 specific cohorts:

Tested – patients tested for SARS-CoV-2 (any test):

- have a record of a first test for SARS-CoV-2 (index event) regardless of result after December 1st 2019. https://atlas.ohdsi.org/#/cohortdefinition/205

Positive test - patients tested positive for SARS-CoV-2;

- have both a record of a test and a record of a positive test for SARS-CoV-2 (index event will be the earliest test date that occurs within 7 days of positive test result) after December 1st, 2019. https://atlas.ohdsi.org/#/cohortdefinition/203

Catch-all - patients with COVID-19 diagnosed OR with a positive test.

- have a record of a test for SARS-CoV-2 (index event first test) after December 1st, 2019, and either
- have a record of a positive test for SARS-CoV-2 OR have a record of COVID-19 diagnosis https://atlas.ohdsi.org/#/cohortdefinition/202

Patients were not required to have a minimum observation period within the database and the study period is 1st December 2019 until the data lock point for each database.

Results

Six out of 8 databases have COVID-19 patients ranging from 1,417 to 124,221 up to June 2020 (data lock points differ) and two of them have less than 5 patients with COVID-19 diagnosis or test (LPD Belgium and IMRD UK). Both IMRD UK and BE datasets are expected to have patients included after the next data update (June 2020 for IMRD UK and July 2020 for LPD Belgium).

Only SIDIAP database contains COVID-19 test data while the all other databases based their diagnosis on medical codes. The patients' numbers and attrition charts for each cohort are found here https://dqdashboard.iqvia.com/ema_report1/ .

7.2 Assessment of variables coverage

The following tables provide an overview of desired variables within each of the 8 databases (Table 3). The list of variables is constructed based on the minimum requirements requested in the Technical specification document, Multicentre collaboration for COVID-19 observational studies, EMA/198302/2020.

Table 3: Presence of desired variables

	LPD Belgium	LPD France	DA Germany	IMRD UK	LPD Italy	IPCI Netherlands	SIDIAP Spain	HM Hospitales Spain
Ability to link data at patient level	✓	✓	✓	✓	✓	✓	✓	✓
Unique patient identifier	✓	✓	✓	✓	✓	✓	✓	✓
Socio-demographic characteristics								
Patient age	✓	✓	✓	✓	✓	✓	✓	✓
Vital status	x	x[1]	x	✓	x[2]	✓	✓	✓
Gender	✓	✓	✓	✓	✓	✓	✓	✓
Ethnicity	x	x	x	x	x	x	x	x
Lifestyle factors								
BMI[3]	✓	✓	✓	✓	✓	✓	✓	x
Socioeconomic status	x	x	x	x	x	Partial	x	x
Smoking[3]	x	x	✓	✓	✓	✓	✓	x
Current and past diagnosis of other medical conditions								
Any before COVID-19 diagnosis	✓	✓	✓	✓	✓	✓	✓	✓
Any after COVID-19 diagnosis	✓	✓	✓	✓	✓	✓	✓	✓
COVID-19 disease history								
Patients initial diagnosis date	✓	✓	✓	✓	✓	✓	✓	✓

	LPD Belgium	**LPD France**	**DA Germany**	**IMRD UK**	**LPD Italy**	**IPCI Netherlands**	**SIDIAP Spain**	**HM Hospitales Spain**
COVID diagnosis	✓	✓	✓	✓	✓	✓	✓	✓
				Laboratory testing				
Presence of any COVID-19 test	x	x	x	x	x	x	✓	x
Type of any COVID-19 test	x	x	x	x	x	x	x	x
Result of any COVID-19 test	x	x	x	x	x	x	✓	x
Other laboratory tests captured[2]	✓	✓	✓	✓	✓	✓	✓	✓
Result of other tests captured[3]	✓	✓	✓	✓	✓	✓	✓	✓
				Interactions with healthcare professionals				
Hospital admission	x	x	x	Partial	x	Partial	✓	✓
Primary care visits	✓	✓	✓	✓	✓	✓	✓	x
				Current and past dug treatment data				
Date of prescription	✓	✓	✓	✓	✓	✓	✓	✓
Type of medication prescribed	✓	✓	✓	✓	✓	✓	✓	✓
In-hospital prescribing	x	x	x	x	x	x	x	✓

1- Possible through linkage with other databases, to be investigated

2- Reported in some instances however low quality

3- BMI can be directly available in the databases or derived from weight and height. These variables are present however sparsely populated and missing not at random, usually being better recorded for patients with specific comorbidities

4- In some databases laboratory tests (measurement) data and the tests values are quite limited, only covering certain tests.

Results

Most of the essential variables are well captured, except for socioeconomic status and ethnicity which are missing in all databases. Also, in three out of eight databases, the vital status is not captured, and this would require linkage with National statistics or other national databases. Hospital admissions are captured in three databases and possible in a fourth one through linkage.

With regards to COVID-19 diagnosis, only one database contains laboratory related data and the type of test is not captured in structured fields at the moment. Two more databases (IMRD UK and LPD Italy) will supplement their records with laboratory tests and more COVID-19 cases at the next database refresh (personal communication from database custodians).

More details on the variable's coverage, where specific variables can be investigated, are found online at https://dqdashboard.iqvia.com/ema_report1/ .

7.3 Assessment of data distribution

Distributions checks were performed to determine whether the skew of data within a variable is logical and broadly as expected. These were performed in the Catch-all cohort of COVID patients.

Distribution checks were performed on specific variables within each dataset which encompassed:

- Socio-demographics
- Vital status
- Hospitalisation distribution (hospitalised and non-hospitalised)
- Severity of COVID-19 defined in relation to the type of oxygen supplementation (no oxygen needed; requiring oxygen in any form; requiring intensive services[1]; requiring extracorporeal membrane oxygenation – ECMO)
- Available follow-up time before and after diagnosis date (min-max, median and interquartile ranges (Q1-Q3)

Results

Information on age and gender is available in all databases. Ethnicity information is not available in any of the databases. Mortality data is available for 5 out of 8 databases, and hospitalisation data is available in three databases. Oxygen supplementation is available in secondary care database (HM Hospitales).

The median follow-up time is 36 days and the median lookback window is 2,302 days. There are significantly shorter follow-up times for the secondary care database, HM Hospitales. The complete results of these distribution checks are shown here https://dqdashboard.iqvia.com/ema_report1/

[1] Intensive services are defined as a record of mechanical ventilation or tracheostomy or ECMO during hospitalisation.

7.4 Ethics approvals and access restriction to the data

Data custodians of the eight included databases answered questionnaire about any access restrictions to the data and ethics approvals manually. Results are shown in the table below.

Table 4: Access restrictions and ethics approval

	LPD Belgium	**LPD France**	**DA Germany**	**IMRD UK**	**LPD Italy**	**IPCI Netherlands**	**SIDIAP Spain**	**HM Hospitales Spain**
Contact of responsible person regarding legal aspects	✓	✓	✓	✓	✓	✓	✓	✓
Contact of responsible person regarding privacy aspects of the data	✓	✓	✓	✓	✓	✓	✓	✓
Institutional review board (IRB) process	No IRB process	No IRB process	No IRB process	6 weeks	No IRB process	Max 1 month, however the IRB meets only 2-4 times a year.	2-3 months	2-4 months
Accept central IRB or require local sign-off	Local sign off	Local sign off	Local sign off	Local sign off	Local sign off	Central IRB not accepted	Each site has to be contracted / submitted individually, or at least per region	Local sign off
Additional ethics review	x	x	x	x	x	✓	✓	✓

Section 8.0 Next steps

This report will be updated every time a new database joins the network, up to June2021.

The following steps are:

- Conduct a proof of concept study on COVID-19 to test the network capabilities
- Describe the collaborative framework including the mechanisms by which the collaboration could be used efficiently by the researchers, governance aspects, data ownership and access, processes for data extraction and analysis and collaborative agreements.
- Create a template(s) of study protocol(s) for the use of the collaboration to conduct multinational pharmacoepidemiological studies related to COVID-19 infection
- Report the results of proof-of concept study
- Analysis of the efficiency of the collaboration between data sources, possible improvements and possible future developments.

It is envisaged that the created network can be used for future COVID-19 research, after this project ends.

References

ECDC. (2020). COVID-19 situation update worldwide *European Centre for Disease Prevention and Control.* Retrieved from https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases

EHDEN. (2020). OHDSI is running a virtual study-a-thon on COVID-19. *European Health Data & Evidence Network*.

Group, R. C., Horby, P., Lim, W. S., Emberson, J. R., Mafham, M., Bell, J. L., . . . Landray, M. J. (2020). Dexamethasone in Hospitalized Patients with Covid-19 - Preliminary Report. *N Engl J Med*. doi:10.1056/NEJMoa2021436

Hall, G. C., Sauer, B., Bourke, A., Brown, J. S., Reynolds, M. W., & LoCasale, R. (2012). Guidelines for good database selection and use in pharmacoepidemiology research. *Pharmacoepidemiol Drug Saf, 21*(1), 1-10. doi:10.1002/pds.2229

Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., . . . Schilling, L. (2016). A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. *EGEMS (Wash DC), 4*(1), 1244. doi:10.13063/2327-9214.1244

Li, J., Wang, X., Chen, J., Zhang, H., & Deng, A. (2020). Association of Renin-Angiotensin System Inhibitors With Severity or Risk of Death in Patients With Hypertension Hospitalized for Coronavirus Disease 2019 (COVID-19) Infection in Wuhan, China. *JAMA Cardiol*. doi:10.1001/jamacardio.2020.1624

Meng, J., Xiao, G., Zhang, J., He, X., Ou, M., Bi, J., . . . Zhang, G. (2020). Renin-angiotensin system inhibitors improve the clinical outcomes of COVID-19 patients with hypertension. *Emerg Microbes Infect, 9*(1), 757-760. doi:10.1080/22221751.2020.1746200

OHDSI. (2020a). Mission, vision and values of OHDSI. *Observational health data sciences and informatics*.

OHDSI. (2020b). OHDSI COVID-19 study-a-thon and evaluation of safety of hydroxychloroquine in RA patients. *Observational health data sciences and informatics*.

Pacurariu, A., Plueschke, K., McGettigan, P., Morales, D. R., Slattery, J., Vogl, D., . . . Cave, A. (2018). Electronic healthcare databases in Europe: descriptive analysis of characteristics and potential for use in medicines regulation. *BMJ Open, 8*(9), e023090. doi:10.1136/bmjopen-2018-023090

Sarzi-Puttini, P., Giorgi, V., Sirotti, S., Marotto, D., Ardizzone, S., Rizzardini, G., . . . Galli, M. (2020). COVID-19, cytokines and immunosuppression: what can we learn from severe acute respiratory syndrome? *Clin Exp Rheumatol, 38*(2), 337-342.

Singh, A. K., Majumdar, S., Singh, R., & Misra, A. (2020). Role of corticosteroid in the management of COVID-19: A systemic review and a Clinician's perspective. *Diabetes Metab Syndr, 14*(5), 971-978. doi:10.1016/j.dsx.2020.06.054

Torjesen, I. (2020). Covid-19: Hydroxychloroquine does not benefit hospitalised patients, UK trial finds. *BMJ, 369*, m2263. doi:10.1136/bmj.m2263

WHO. (2020). Coronavirus disease (COVID-19) pandemic. *World Health Organization*.

Zhang, P., Zhu, L., Cai, J., Lei, F., Qin, J. J., Xie, J., . . . Li, H. (2020). Association of Inpatient Use of Angiotensin-Converting Enzyme Inhibitors and Angiotensin II Receptor Blockers With Mortality Among Patients With Hypertension Hospitalized With COVID-19. *Circ Res, 126*(12), 1671-1681. doi:10.1161/CIRCRESAHA.120.317134

Appendix 1 Log of errors found during quantitative assessment

Date	Description of the issue	Status (Solved/Investigating/Not solved)	Details
10th September 2020	Follow-up time appears longer than it should be based on data lock point for IQVIA databases	Under investigation	in the absence of enrollment dates in the source data the OMOP *observation_period_end_date* is calculated as the max of patient event dates. Some event dates are recorded as being in the future.