Title

Characterising COVID-19 related events in a nationwide electronic health record cohort of 55.9 million people in England

Authors

Spiros Denaxas, Christopher Tomlinson, Johan Thygesen

Version history

Version	Date	Notes
0.1	01/06/2021	First version
0.11	25/06/2021	First version to GitHub

Lay summary

When a patient visits their GP or is admitted into hospital, information about their symptoms, diagnosis, lab test results and prescriptions is inputted and stored in ‘Electronic Health Records’ (‘EHRs’). These EHR’s are a valuable resource for researchers and clinicians to be able to analyse the health data of large numbers of patients, with the aim of using this information to improve patient health and care.

However, as information in these EHRs is inputted by different health workers, and different hospitals, and GP practices use different EHR systems, the records can look very different whilst sharing the same underlying meaning. This means that researchers need to initially spend a considerable amount of time and effort to obtain the most relevant information from these EHRs, before they can begin to effectively analyse them. COVID-19 has added to the difficulty of this, as new methods of recording infection and testing had to be created and changed to keep pace with the pandemic. This makes seemingly simple tasks complex, such as determining which patients had COVID and what happened to them, whether they were admitted, received a certain type of ventilation, or died.

The approaches developed in this project will be shared with both researchers in the CVD-COVID-UK consortium and the wider scientific and medical community by publishing the results openly. This will maximise the benefits of using information from Electronic Health Records, and ensure research can be reproduced effectively. Most importantly, this will speed up the ability to effectively analyse health information in EHRs, answer vital questions and directly benefit patients and healthcare.

Background

Currently, several large UK studies have explored factors associated with COVID-19 outcomes using linked electronic health record (EHR) data. Such studies are based on sub-population cohorts, limited data sources and report coarse outcomes such as mortality and hospitalisation.

To the best of our knowledge there has to date not been a national study exploring COVID-19 infection and severity with population-scale EHR and administrative data spanning multiple healthcare settings and with extensive description of ventilatory support and disease trajectories. This project seeks to address this by comprehensively characterising individual COVID-19 related events and disease trajectories to create and evaluate disease severity phenotypes using data on a population-wide scale.

Research aims and hypothesis

This project aims to identify and describe patterns of recording of COVID-19 infection and related events across multiple linked national electronic health record data sources in the CVD-COVID-UK NHS England TRE. Specifically, this project will explore the following questions:

1. How many people have had a COVID-19 infection event recorded in any data source and how does recording vary by pandemic wave?
2. What is the concordance of recording COVID-19 infection events across data sources and do any temporal differences exist?
3. What are the trajectories of infection amongst people with COVID-19 infection and how do the trajectories vary across pandemic waves?

Data sources

NHS Digital TRE for England

- Primary care data - General Practice Extraction Service Extract for Pandemic Planning and Research (GDPPR)
- Secondary care data - Secondary Use Service (SUS), Hospital Episode Statistics (HES) Admitted Patient Care (APC), Outpatient (OP) & Critical Care (CC), COVID-19 Hospitalisation in England Surveillance System (CHESS)
- COVID testing data - Pillar 1 and Pillar 2 SARS-CoV-2 infection laboratory testing data from Second Generation Surveillance System (SGSS)
- Office of National Statistics death registration records

Study design

Cohort study using linked electronic health records.

The study start date is defined as 23/01/2020 and the study end date is 31/03/2021.

Study population

Patients are included if they meet all of the following criteria:

- Valid and non-missing patient pseudoidentifier
- Alive and registered with a GP practice on the study start date
- Located in England as defined by their LSOA
- Minimum 28 days of follow up time

- One or more COVID-19 related events (defined in the Exposure section below) during the study period.

Patients are excluded if they meet any of the following criteria:

- Missing date of birth information
- Missing sex

Individuals enter the study on the date of the earliest COVID-19-related event in any source and are censored at the earliest of the date of death or the study end date.

Exposure

A rule-based phenotyping algorithm will be developed to ascertain infection status and onset across all sources using a combination of controlled clinical terminology terms (e.g. SNOMED-CT concepts for GDPPR and ICD-10 codes in HES APC) or source-specific criteria (e.g. SGSS).

Outcomes

Primary outcomes are hospital admissions with COVID-19, ventilatory support (encompassing Non-Invasive Ventilation (NIV), Invasive Mechanical Ventilation (IMV), Intensive Care Unit (ICU) admission and Extracorporeal Membrane Oxygenation (ECMO)) and death. COVID-19 admissions were defined as anyone with a hospital admission in CHESS or an admission with a COVID-19 diagnosis in HES APC or SUS, in any position.

Provision of ventilatory support will be defined from multiple sources: a) CHESS, b) HES CC, d) SUS, e) HES APC.

Fatal COVID-19 events will be identified from ONS mortality data and classified based on the presence of COVID-19 ICD-10 codes on the death certificate and time to death since infection.

Covariates

Covariates will be defined for all study participants from multiple linked sources (GDPPR, HES, ONS):

- Age
- Sex
- Ethnicity
- Socioeconomic deprivation information (Index of Multiple Deprivation)

Statistical analyses

Descriptive statistics including means, median and proportions will be used to summarize patient populations and subpopulations. Chi-squared test will be used to test for significance between subpopulation covariates.

We will plot COVID-19 trajectory networks based on individual trajectories, identified as the date ordered progression of any of the following events; diagnosis, positive test, hospital admission, NIV treatment, ICU admission, IMV treatment, ECMO treatment and death with COVID-19 diagnosis on death register or COVID-19 death without diagnosis on death register.

We will calculate consecutive positive tests and evaluate different timing windows between tests in order to define reinfection (initial value will be >=90 days)

Key planned outputs

1. Manuscript detailing phenotyping methodology, sub-population characteristics and descriptive statistics. To be published in consortium name and available Open Access.
2. Public code repository sharing codelists and scripts to allow reproducibility of phenotyping approaches on external datasets.
3. Internal tables including COVID-19 events and a 'research-ready' to facilitate efficient collaboration between consortium members within the NHS Digital TRE

Target Tables & Figures

Table 1: Characteristics of people with a confirmed or suspected COVID-19 diagnosis, stratified by overall COVID-19 severity phenotypes. Values are numbers (percentages) unless otherwise specified.Ethnicity information is derived from both primary care and hospitalization records. Patients at high risk for developing complications from infection were identified from primary care using the NHS Digital SNOMED flag term. Comorbidities (Stroke/Transient Ischaemic Attack (TIA), Myocardial Infarction (MI), Diabetes & Obesity) ascertained from both Primary & Secondary care records. * indicates p < 0.001

	All COVID cases	**Hospitalisation**		**Ventilatory Support**		**COVID Fatalities**	
		No	**Yes**	**No**	**Yes**	**No**	**Yes**
n (%)							
Male (%)							
Age bracket (%)							
Under 18							
18 - 29							
30 - 49							
50 - 69							
Over 70							
Ethnicity (%)							
White							
Asian or Asian British							
Black or Black British							
Chinese							
Mixed							

Other							
Unknown							
IMD fifths (%)							
1 (most deprived)							
5 (least deprived)							
High Risk (%)							
Long COVID diagnosis(%)							
Prev. Stroke/TIA (%)							
Prev. MI (%)							
Prev. Diabetes (%)							
Prev. Obesity (%)							

Target Figures

Figure 1: Flowchart of cohort design showing the number of records/individuals (*n*) excluded at different stages and the identification of cases and the final study population.

Figure 2: Flowchart of phenotyping process used to ascertain COVID-19 events in cohort. Information derived from the following data sources; SGSS (Second Generation Surveillance System) Pillars 1 & 2, Pillar 2 antigen, primary care EHR from GDPPR, hospitalisation EHR from HES Admitted Patient Care, Critical Care and Outpatient, SUS (Secondary Uses Service) and CHESS (COVID-19 Hospitalisations in England Surveillance System), national death registrations from the ONS. Sources used to identify each step are indicated with data buckets on the left and COVID-related events in rectangles on the right. Ventilation support is defined either as Non-Invasive Ventilation (NIV), Invasive Mechanical Ventilation (IMV) or Extracorporeal Membrane Oxygenation (ECMO). HES CC does not give info on ECMO treatments. Fatal COVID-19 events are defined as events at any point in time with COVID-19 recorded as the cause of death (at any position on the death certificate) or within 28-days of the earliest COVID-19 ascertainment event irrespective of the cause of death recorded on the death certificate. In all sources, ontology terms for both suspected and confirmed diagnosis were used.

Figure 3: venn Diagram of data sources reporting person level data on confirmed or suspected COVID-19 diagnoses between 1st January 2020 and 1st May 2021. Numbers indicate subgroup sizes of distinct individuals with COVID-19 (suspected or confirmed diagnosis). Primary Care data derived from GDPPR, COVID-19 Testing from SGSS Pillar 1 & 2 and Pillar 2 antigen testing, Deaths from ONS deaths registry and Hospital Episodes from HES APC, CC, CHESS and SUS.

Figure 4: Timeline of COVID-19 events. Note this shows unique events per individual per date, a person may have multiple events of the same type at different dates. (COVID-19 lab test positive, COVID-19 diagnosis in primary care, COVID-19 hospital admission, ICU admission, NIV, IMV & ECMO treatments, deaths with COVID-19 diagnosis on death certificate, deaths within 28 days of a positive test without COVID diagnosis on death certificate.

Figure 5: COVID-19 trajectory plots. Panel (A) includes all events, panel (B) focuses only on events after hospital admission. The size of the circles and width of the paths are relative to the number of patients with that event or transition. Percentages are shown for paths with >= 1% of the total transitions. Note that individuals may have

anywhere from only one event recorded up to a maximum of 8 events. See methods for further detail on how trajectories were identified.

Appendix: COVID-19 phenotype codelists

COVID Phenotype	Code	Terminology	Description	Source
01_Covid_diagnosis	1321301000000101	SNOMED	Severe acute respiratory syndrome coronavirus 2 ribonucleic acid qualitative existence in specimen (observable entity)	GDPPR
01_Covid_diagnosis	1321761000000103	SNOMED	Severe acute respiratory syndrome coronavirus 2 immunoglobulin A detected (finding)	GDPPR
01_Covid_diagnosis	1321541000000108	SNOMED	Severe acute respiratory syndrome coronavirus 2 immunoglobulin G detected (finding)	GDPPR
01_Covid_diagnosis	1240561000000108	SNOMED	Encephalopathy caused by severe acute respiratory syndrome coronavirus 2 (disorder)	GDPPR
01_Covid_diagnosis	1240401000000105	SNOMED	Antibody to severe acute respiratory syndrome coronavirus 2 (substance)	GDPPR
01_Covid_diagnosis	1322871000000109	SNOMED	Severe acute respiratory syndrome coronavirus 2 antibody detection result positive (finding)	GDPPR
01_Covid_diagnosis	1300631000000101	SNOMED	Coronavirus disease 19 severity score (observable entity)	GDPPR
01_Covid_diagnosis	1300721000000109	SNOMED	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 confirmed by laboratory test (situation)	GDPPR
01_Covid_diagnosis	186747009	SNOMED	Coronavirus infection (disorder)	GDPPR
01_Covid_diagnosis	1321811000000105	SNOMED	Severe acute respiratory syndrome coronavirus 2 immunoglobulin A qualitative existence in specimen (observable entity)	GDPPR
01_Covid_diagnosis	120814005	SNOMED	Coronavirus antibody (substance)	GDPPR
01_Covid_diagnosis	U07.1	ICD10	Confirmed_COVID19	HES OP
01_Covid_diagnosis	1240541000000107	SNOMED	Infection of upper respiratory tract caused by severe acute respiratory syndrome coronavirus 2 (disorder)	GDPPR
01_Covid_diagnosis	1321201000000107	SNOMED	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 health issues simple reference set (foundation metadata concept)	GDPPR
01_Covid_diagnosis	1300681000000102	SNOMED	Assessment using coronavirus disease 19 severity scale (procedure)	GDPPR
01_Covid_diagnosis	1029481000000103	SNOMED	Coronavirus nucleic acid detection assay (observable entity)	GDPPR
01_Covid_diagnosis	1240421000000101	SNOMED	Serotype severe acute respiratory syndrome coronavirus 2 (qualifier value)	GDPPR
01_Covid_diagnosis	1008541000000105	SNOMED	Coronavirus ribonucleic acid detection assay (observable entity)	GDPPR
01_Covid_diagnosis	1240571000000101	SNOMED	Gastroenteritis caused by severe acute respiratory syndrome coronavirus 2 (disorder)	GDPPR
01_Covid_diagnosis	1321241000000105	SNOMED	Cardiomyopathy caused by severe acute respiratory syndrome coronavirus 2 (disorder)	GDPPR
01_Covid_diagnosis	1300731000000106	SNOMED	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 confirmed using clinical diagnostic criteria (situation)	GDPPR
01_Covid_diagnosis	1322781000000102	SNOMED	Severe acute respiratory syndrome coronavirus 2 antigen detection result positive (finding)	GDPPR
01_Covid_diagnosis	1240581000000104	SNOMED	Severe acute respiratory syndrome coronavirus 2 ribonucleic acid detected (finding)	GDPPR
01_Covid_diagnosis	1240521000000100	SNOMED	Otitis media caused by severe acute respiratory syndrome coronavirus 2 (disorder)	GDPPR
01_Covid_diagnosis	1300671000000104	SNOMED	Coronavirus disease 19 severity scale (assessment scale)	GDPPR
01_Covid_diagnosis	1321351000000100	SNOMED	Arbitrary concentration of severe acute respiratory syndrome coronavirus 2 immunoglobulin M in serum (observable entity)	GDPPR
01_Covid_diagnosis	1240381000000105	SNOMED	Severe acute respiratory syndrome coronavirus 2 (organism)	GDPPR

01_Covid_diagnosis	1321181000000108	SNOMED	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 record extraction simple reference set (foundation metadata concept)	GDPPR
01_Covid_diagnosis	1240751000000100	SNOMED	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 (disorder)	GDPPR
01_Covid_diagnosis	1321551000000106	SNOMED	Severe acute respiratory syndrome coronavirus 2 immunoglobulin M detected (finding)	GDPPR
01_Covid_diagnosis	1240511000000106	SNOMED	Detection of severe acute respiratory syndrome coronavirus 2 using polymerase chain reaction technique (procedure)	GDPPR
01_Covid_diagnosis	U07.2	ICD10	Suspected_COVID19	HES OP
01_Covid_diagnosis	1321341000000103	SNOMED	Arbitrary concentration of severe acute respiratory syndrome coronavirus 2 immunoglobulin G in serum (observable entity)	GDPPR
01_Covid_diagnosis	1240391000000107	SNOMED	Antigen of severe acute respiratory syndrome coronavirus 2 (substance)	GDPPR
01_Covid_diagnosis	1321321000000105	SNOMED	Severe acute respiratory syndrome coronavirus 2 immunoglobulin G qualitative existence in specimen (observable entity)	GDPPR
01_Covid_diagnosis	1240531000000103	SNOMED	Myocarditis caused by severe acute respiratory syndrome coronavirus 2 (disorder)	GDPPR
01_Covid_diagnosis	1240551000000105	SNOMED	Pneumonia caused by severe acute respiratory syndrome coronavirus 2 (disorder)	GDPPR
01_Covid_diagnosis	1321311000000104	SNOMED	Severe acute respiratory syndrome coronavirus 2 immunoglobulin M qualitative existence in specimen (observable entity)	GDPPR
01_Covid_diagnosis	1321331000000107	SNOMED	Arbitrary concentration of severe acute respiratory syndrome coronavirus 2 total immunoglobulin in serum (observable entity)	GDPPR
01_Covid_diagnosis	1240741000000103	SNOMED	Severe acute respiratory syndrome coronavirus 2 serology (observable entity)	GDPPR
01_Covid_diagnosis	1321191000000105	SNOMED	Coronavirus disease 19 caused by severe acute respiratory syndrome coronavirus 2 procedures simple reference set (foundation metadata concept)	GDPPR
01_Covid_diagnosis	1321801000000108	SNOMED	Arbitrary concentration of severe acute respiratory syndrome coronavirus 2 immunoglobulin A in serum (observable entity)	GDPPR
01_Covid_positive			pillar_1	SGSS
01_Covid_positive			pillar_2	SGSS
01_Covid_positive			Pillar 2	Pillar 2
02_Covid_admission	U07.1	ICD10	Confirmed_COVID19	SUS
02_Covid_admission	U07.2	ICD10	Suspected_COVID19	SUS
02_Covid_admission	U07.1	ICD10	Confirmed_COVID19	HES APC
02_Covid_admission	U07.2	ICD10	Suspected_COVID19	HES APC
02_Covid_admission				CHESS
03_ECMO_treatment	X58.1	OPCS	Extracorporeal membrane oxygenation	SUS
03_ECMO_treatment	X58.1	OPCS	Extracorporeal membrane oxygenation	HES APC
03_ECMO_treatment			RespiratorySupportECMO == Yes	CHESS
03_ICU_admission			id is in hes_cc table	HES CC
03_ICU_admission				CHESS
03_IMV_treatment			Invasivemechanicalventilation == Yes	CHESS
03_IMV_treatment	X56	OPCS	Intubation of trachea	SUS
03_IMV_treatment	X56	OPCS	Intubation of trachea	HES APC
03_IMV_treatment			ARESSUPDAYS > 0	HES CC
03_IMV_treatment	E85.1	OPCS	Invasive ventilation	SUS
03_IMV_treatment	E85.1	OPCS	Invasive ventilation	HES APC
03_NIV_treatment			Highflownasaloxygen OR NoninvasiveMechanicalventilation == Yes	CHESS
03_NIV_treatment	E85.6	OPCS	Continuous positive airway pressure	SUS

03_NIV_treatment	E85.6	OPCS	Continuous positive airway pressure	HES APC
03_NIV_treatment			bressupdays > 0	HES CC
03_NIV_treatment	E85.2	OPCS	Non-invasive ventilation NEC	SUS
03_NIV_treatment	E85.2	OPCS	Non-invasive ventilation NEC	HES APC
04_Fatal_with_covid_diagnosis				deaths
04_Fatal_without_covid_diagnosis			ONS death within 28 days	deaths