eMethods

TriNetX network

This section reproduces our previous description of the network.[1]

Legal and ethical status
TriNetX's Analytics network is compliant with the Health Insurance Portability and Accountability Act (HIPAA), the US federal law which protects the privacy and security of healthcare data. TriNetX is certified to the ISO 27001:2013 standard and maintains an Information Security Management System (ISMS) to ensure the protection of the healthcare data it has access to and to meet the requirements of the HIPAA Security Rule. Any data displayed on the TriNetX Platform in aggregate form, or any patient level data provided in a data set generated by the TriNetX Platform, only contains de-identified data as per the de-identification standard defined in Section §164.514(a) of the HIPAA Privacy Rule. The process by which the data is de-identified is attested to through a formal determination by a qualified expert as defined in Section §164.514(b)(1) of the HIPAA Privacy Rule. This formal determination by a qualified expert, refreshed in December 2020, supersedes the need for TriNetX's previous waiver from the Western Institutional Review Board (IRB). The network contains data that are provided by participating Health Care Organizations (HCOs), each of which represents and warrants that it has all necessary rights, consents, approvals and authority to provide the data to TriNetX under a Business Associate Agreement (BAA), so long as their name remains anonymous as a data source and their data are utilized for research purposes. The data shared through the TriNetX Platform are attenuated to ensure that they do not include sufficient information to facilitate the determination of which HCO contributed which specific information about a patient.

Acquisition of data, quality control, and other procedures
The data are stored onboard a TriNetX appliance – a physical server residing at the institution's data centre or a virtual hosted appliance. The TriNetX platform is a fleet of these appliances connected into a federated network able to broadcast queries to each appliance. Results are subsequently collected and aggregated.

Once the data are sent to the network, they are mapped to a standard and controlled set of clinical terminologies and undergo a data quality assessment including 'data cleaning' that rejects records which do not meet the TriNetX quality standards. HIPAA compliance of the clinical patient data is achieved using de-identification. Different data modalities are available in the network. They include demographics (coded to HL7 version 3 administrative standards), diagnoses (represented by ICD-10-CM codes), procedures (coded in ICD-10-PCS or CPT), measurements (coded to LOINC), and clinical drugs (represented as VA class and/or RxNorm). While extensive information is provided about patients' diagnoses and procedures, other variables (such as socioeconomic and lifetime factors are not comprehensively represented).

The data from a typical HCO generally go back around 7 years, with some going back 13 years. The data are continuously updated. HCOs update their data at various times, with most refreshing every 1, 2, or 4 weeks.

The data come primarily (>93%) from HCOs in the USA, with the remainder coming from India, Australia, Malaysia, Taiwan, Spain, UK, and Bulgaria. As noted above, to comply with legal frameworks and ethical guidelines guarding against data re-identification, the identity of participating HCOs and their individual contribution to each dataset are not disclosed to researchers.

Data quality assessment followed a standardised strategy wherein the data are reviewed for conformance (adherence to specified standards and formats), completeness (quantifying data presence or absence) and plausibility (believability of the data from a clinical perspective). There are pre-defined metrics for each of the above assessment categories. Results for these metrics are visualised and reviewed for each new site that joins the network as well as on an ongoing basis. Any identified issue is communicated to the data provider and resolved before continuing data collection.

The basic formatting of contributed data is also checked (e.g. to ensure that dates are properly represented). Records are checked against a list of required fields (e.g., patient identifier) and rejects those records for which the required information is missing. Referential integrity checking is done to ensure that data spanning multiple database tables can be successfully joined together. As the data are refreshed, changes in volume of data over

time is monitored to ensure data validity. At least one non-demographic fact for each patient is required for them to be counted in the dataset. Patient records with only demographics information are discarded.

The software also undergoes quality control. The engineers testing the software are independent from the engineers developing it. Each test code is checked by two independent testing engineers. Each piece of software is tested extensively against a range of synthetic data (i.e. generated for the purpose of testing) for which the expected output is established independently. If the software fails to return this output, then the software is deemed to have failed the test and is examined and modified accordingly. For statistical software (including that used for propensity score matching, for Kaplan-Meier analysis, etc), an additional quality control step is implemented. Two independent codes are written in two different programming languages (typically R and python) and the statistical results are compared. If discrepancies are identified, then the codes are deemed to have failed the test and are examined and modified accordingly. All the code is reviewed independently by another engineer.

The test strategy follows three levels of granularity:

1. Unit tests: These test specific blocks, or units, of code that perform specific actions (e.g. querying the database).
2. Integration tests: These ensure that different components are working together correctly.
3. End-to-end tests: These tests run the entire system and check the final output.

Some comments on advantages and disadvantages of EHR data
One advantage of EHR data, like those in TriNetX, over insurance claim data is that both insured and uninsured patients are included. An advantage of EHR data over survey data is that they represent the diagnostic rates in the population presenting to healthcare facilities. This provides an accurate account of the burden of specific diagnoses on healthcare systems. However, there are also limitations inherent to research using electronic health records,[2–4] including TriNetX:

1. Patients with acute or post-acute sequelae of COVID-19 but were not diagnosed are not included leading to underestimation of actual incidences.
2. Despite the matching and use of various comparison cohorts, there may well be residual confounding, particularly related to social and economic factors which are not well captured in EHR networks and which might influence outcomes post COVID-19.
3. We do not know which diagnoses were made in primary or secondary care or specialist facilities, nor by whom.
4. A patient may be seen in different HCOs for different parts of their care, and if one HCO is not part of the federated network then part of their medical records may not be available. Using a network of HCOs (rather than a single HCO) limits this possibility but does not fully remove it.
5. Since the data are presented as they are recorded, we cannot be sure that there has not been mis-recording of information, adding a degree of noise to the data.
6. Historical data before the start of EHRs (or the addition of an HCO to the network) may well be incomplete.

Definition of cohorts

The control cohort used consisted of patients with a diagnosis of influenza. Specifically, patients with influenza were those who had any of the following diagnoses:
- J09: Influenza due to certain identified influenza viruses
- J10: Influenza due to other identified influenza virus
- J11: Influenza due to unidentified influenza virus.

Because some patients with the control index event might have had COVID-19 at a different point in time, we excluded from the control cohorts all those who had COVID-19 at any point in time. To avoid any contamination between cohorts, COVID-19 as an exclusion criterion was defined in the broader sense to be all patients with a confirmed diagnosis of COVID-19 (ICD-10 code U07.1) but also patients with an unconfirmed COVID-19 diagnosis (U07.2), a recorded positive PCR test for COVID-19, or any of the following recorded on or after January 20, 2020: Pneumonia due to SARS-associated coronavirus (J12.81), Other coronavirus as the cause of disease classified elsewhere (B97.29), or Coronavirus infection unspecified (B34.2). Inclusion of the latter three diagnostic codes captures patients who receive a COVID-19 diagnosis in the early stage of the

pandemic when the ICD code for COVID-19 (U07) was not yet defined. Specifically, the following codes were excluded from the control cohort if they occurred on or after January 20, 2020:

- U07.1: COVID-19, virus identified
- U07.2: COVID-19, virus not identified
- J12.81: Pneumonia due to SARS-associated coronavirus
- B97.29: Other coronavirus as the cause of disease classified elsewhere
- B34.2: Coronavirus infection, unspecified
- Positive SARS-CoV-2 RNA in Respiratory specimen
- Positive SARS-CoV-2 RNA in Unspecified specimen
- Positive SARS-CoV-2 N gene in Respiratory specimen
- Positive SARS-CoV-2 N gene in Unspecified specimen
- Positive SARS-CoV-2 RdRp gene in Respiratory specimen
- Positive SARS-CoV-2 E gene in Respiratory specimen
- Positive SARS-CoV-2 E gene in Unspecified specimen
- Positive SARS-CoV-2 RNA panel in Respiratory specimen
- Positive SARS-CoV-2 RNA panel in Unspecified specimen
- Positive SARS-CoV-2 RNA in Nasopharynx
- Positive SARS coronavirus 2 and related RNA
- Positive SARS-related coronavirus RNA in Respiratory specimen
- Positive SARS coronavirus 2 ORF1ab in Respiratory specimen

The duration of follow-up of the patients depended on when they had the index event. Patients who had the index events more than 6 months before the date of the analysis (August 24, 2021) had 6 months of follow-up. The other patients were followed up until August 24, 2021. The Kaplan-Meier estimator accommodates differences in duration of follow-up by means of censoring.

To keep the cohorts as homogeneous as possible (and because sample size was not an issue), we included fewer patients in the cohort of interest (by only including those with a diagnostic code for COVID-19) and excluded more patients from the control cohort (by excluding both those with a diagnostic code and those with a positive tests). Making the cohorts more homogeneous decreases the sensitivity of the findings to bias even when it comes at the price of a smaller sample size (for an excellent discussion of that point, see Rosenbaum P. Observation and experiment. Chapter 10. Harvard University Press).

Definition of covariates

To reduce the effect of confounding on associations, cohorts were matched for established or suspected risk factors for COVID-19[5–8] and for established risk factors for COVID-19 death[9] (taken to be risk factors of a more severe COVID-19 illness). These were the covariates used in our previous studies.[1,10,11] The following confounding factors were therefore included (with ICD-10/VA Class/RxNorm codes in brackets):

1) **Age** at the time of diagnosis.
2) **Sex** coded as female, male, or other.
3) **Race** encoded as 6 separate dichotomous variables: White (2106-3), Black or African American (2054-5), American Indian or Alaska Native (1002-5), Asian (2028-9), Native Hawaiian or Other Pacific Islander (2076-8), or Unknown Race (2131-1).
4) **Ethnicity** encoded as Hispanic or Latino (2135-2), Not Hispanic or Latino (2186-5), or Unknown Ethnicity.
5) **Socioeconomic deprivation** encoded as the ICD-10 code for Problems related to housing and economic circumstances (Z59).
6) **Obesity** encoded as one dichotomous variable and one categorical variable: Overweight and obesity (E66) and body mass index (categorised into < 25 kg/m^2, 25-30 kg/m^2, ≥ 30 kg/m^2 which are the WHO thresholds for not obese, pre-obese, and obesity).
7) **Hypertension** encoded as 2 dichotomous and 2 categorical variables: Hypertensive diseases (I10-I16), the now deprecated version that was used until 2018 Hypertension diseases (I10-I15), measurements of systolic blood pressure (categorised into < 140mmHg, 140-160mmHg, and ≥ 160mmHg), and diastolic blood pressure (categorised into < 90mmHg, 90-100mmHg, and ≥ 100mmHg). The blood pressure categories correspond to the absence of hypertension, stage 1 hypertension, and stage 2 (and over) hypertension as per the NICE guidelines.

8) **Diabetes mellitus** encoded as 2 dichotomous variables: Type 1 diabetes mellitus (E10) and Type 2 diabetes mellitus (E11).
9) **Chronic lower respiratory diseases** encoded by each sub-category of the corresponding ICD-10 group: Bronchitis, not specified as acute or chronic (J40), Simple and mucopurulent chronic bronchitis (J41), Unspecified chronic bronchitis (J42), Emphysema (J43), Other chronic obstructive pulmonary disease (J44), Asthma (J45), Bronchiectasis (J47).
10) **Nicotine dependence** encoded as the corresponding ICD-10 diagnosis (F17.2).
11) **Substance use disorders** encoded as the ICD-10 code for mental and behavioural disorders due to psychoactive substance use (F10-F19).
12) **Psychotic disorders** encoded as the ICD-10 code for schizophrenia, schizotypal, delusional, and other non-mood psychotic disorders (F20-F29).
13) **Mood disorders** encoded as a single variable (as well as individual codes, see below) with any of the ICD-10 code for mood disorders (F30-F39).
14) **Anxiety disorders** encoded as the ICD-10 code for anxiety, dissociative, stress-related, somatoform and other nonpsychotic mental disorders (F40-F48)
15) **Heart diseases** encoded as 2 categorical variables: Ischaemic heart disease (I20-I25) and Other forms of heart disease (I30-I52).
16) **Chronic kidney disease** encoded as 2 dichotomous variables: Chronic kidney disease (N18) and Hypertensive chronic kidney disease (I12).
17) **Chronic liver disease** encoded as 8 categorical variables: Alcoholic liver disease (K70), Hepatic failure, not elsewhere classified (K72), Chronic hepatitis, not elsewhere classified (K73), Fibrosis and cirrhosis of liver (K74), Fatty (change of) liver, not elsewhere classified (K76.0), Chronic passive congestion of liver (K76.1), Portal hypertension (K76.6), Other specified diseases of liver (K76.8).
18) **Stroke** encoded as the dichotomous variable Cerebral infarction (I63) .
19) **Dementia** encoded as 6 dichotomous variables: Vascular dementia (F01), Dementia in other diseases classified elsewhere (F02), Unspecified dementia (F03), Alzheimer's disease (G30), Frontotemporal dementia (G31.0), and Dementia with Lewy bodies (G31.83).
20) **Cancer and haematological cancer in particular** encoded as 2 dichotomous variables: Neoplasms (C00-D49) and Malignant neoplasms of lymphoid, hematopoietic and related tissue (C81-C96).
21) **Organ transplant** encoded as 2 dichotomous variables: Renal Transplantation Procedures and Liver Transplantation Procedures.
22) **Rheumatoid arthritis** encoded as 2 dichotomous variables: Rheumatoid arthritis with rheumatoid factor (M05) and Other rheumatoid arthritis (M06).
23) **Lupus** encoded as a dichotomous variable corresponding ICD-10 code (M32).
24) **Psoriasis** encoded as a dichotomous variable corresponding ICD-10 code (L40).
25) **Disorders involving an immune mechanism** encoded as a dichotomous variable "Certain disorders involving the immune mechanism" (D80-D89).

Each individual code was considered a confounding factor in and of itself so that matching was achieved for each of them individually. For instance, matching was achieved for each subcategory (and not just for the whole category) of chronic lower respiratory diseases. For variables representing diagnoses and socioeconomic deprivation, an individual was considered positive if the diagnostic was recorded at least once in their health record before the index event. For categorical variables representing measurements (i.e. BMI and blood pressures), all available measurements for all individuals were used and propensity score matching sought to define cohorts with similar numbers of measurements falling into each category.

Details on statistical analyses

Implementation details of propensity score matching
In propensity score matching, the propensity score was calculated using a logistic regression (implemented by the function LogisticRegression of the scikit-learn package in Python 3.7) including each of the covariates mentioned above. To eliminate the influence of ordering of records, the order of the records in the covariate matrix were randomised before matching.

Testing proportional hazards
The assumption that the hazards were proportional when accounting for the two phases was tested using the generalized Schoenfeld approach[12] implemented in the cox.zph function of the survival package (version 3.2.3) in R.

Time-varying hazard ratio
If the proportional hazard assumption was found to be violated (i.e. statistical evidence from a score test indicating a non-zero slope in the scaled Schoenfeld residuals over time), then the time-varying HR was assessed using natural cubic splines (in log-time) to the log-cumulative hazard[13]. This was achieved using the generalized survival models of the rstpm2 package (version 1.5.1) in R[14]. As recommended by Royston and Parmar[13], splines with 1, 2, and 3 degrees of freedom were estimated for both the baseline log-cumulative hazard and its cohort dependency and the number of degrees of freedom leading to the lowest Akaike Information Criterion (AIC) was selected. This was achieved on a per-comparison basis so that more complex time dependency (i.e. higher number of degrees of freedom) could be selected for a specific comparison if there was enough evidence in the data to support such complexity.

Assessing moderation of associations
The Cox model used in the primary analysis reads:

$$\lambda(t|\mathrm{COVID}) = \lambda_0(t)\exp(\beta_0 + \beta_C \mathrm{COVID}),$$

where COVID=1 if the individual had COVID, and 0 if they had influenza. To assess whether associations between COVID status and outcomes are moderated by a variable *X*, we added an interaction term as follows:

$$\lambda(t|\mathrm{COVID}, X) = \lambda_0(t)\exp(\beta_0 + \beta_C \mathrm{COVID} + \beta_X X + \beta_{XC} X \times \mathrm{COVID}).$$

The null hypothesis that $\beta_{XC} = 0$ corresponds to no moderation by *X*. This is tested as part of the estimation of the Cox model with coxph function from the Survival package (version 3.2.3).

In our two secondary analyses, *X* represented age (≤ 16 vs. > 16) and hospitalisation respectively.

Supplementary Tables

eTable 1 – Baseline characteristics for COVID-19 and influenza cohorts before and after matching.

	Before matching			After matching		
	COVID-19	**Influenza**	**SMD**	**COVID-19**	**Influenza**	**SMD**
Number	681283	179651	-	152754	152754	-
DEMOGRAPHICS						
Age; mean (SD); y	45.6 (21.1)	26.5 (22.6)	0.9	31.0 (21.2)	30.2 (22.6)	0.04
Sex; n (%)						
Female	377061 (55.3)	97501 (54.3)	0.02	84111 (55.1)	85077 (55.7)	0.01
Male	303898 (44.6)	82132 (45.7)	0.02	68597 (44.9)	67660 (44.3)	0.01
Other	324 (0.05)	18 (0.01)	0.02	46 (0.03)	17 (0.01)	0.01
Race; n (%)						
White	416830 (61.2)	115596 (64.3)	0.07	99667 (65.2)	97138 (63.6)	0.03
Black or African American	116220 (17.1)	30376 (16.9)	0.004	25631 (16.8)	25390 (16.6)	0.004
Asian	21093 (3.1)	6883 (3.8)	0.04	5370 (3.5)	5227 (3.4)	0.005
American Indian or Alaska Native	2759 (0.4)	750 (0.4)	0.002	562 (0.4)	628 (0.4)	0.007
Native Hawaiian or Other Pacific Islander	1600 (0.2)	345 (0.2)	0.009	286 (0.2)	314 (0.2)	0.004
Unknown	122781 (18.0)	25701 (14.3)	0.1	21238 (13.9)	24057 (15.7)	0.05
Ethnicity; n (%)						
Hispanic or Latino	96390 (14.1)	23166 (12.9)	0.04	19090 (12.5)	21062 (13.8)	0.04
Not Hispanic of Latino	424617 (62.3)	127377 (70.9)	0.2	108005 (70.7)	103628 (67.8)	0.06
Unknown	160276 (23.5)	29108 (16.2)	0.2	25659 (16.8)	28064 (18.4)	0.04
Socioeconomic deprivation; n (%)	6117 (0.9)	987 (0.5)	0.04	930 (0.6)	962 (0.6)	0.003
COMORBIDITIES; n (%)						
Overweight and obesity	139368 (20.5)	23588 (13.1)	0.2	23042 (15.1)	22940 (15.0)	0.002
Hypertensive disease	219503 (32.2)	29479 (16.4)	0.4	31552 (20.7)	29072 (19.0)	0.04
Diabetes mellitus						
Type 1 diabetes mellitus	13778 (2.0)	2103 (1.2)	0.07	2101 (1.4)	2045 (1.3)	0.003
Type 2 diabetes mellitus	109869 (16.1)	12389 (6.9)	0.3	13665 (8.9)	12330 (8.1)	0.03
Chronic lower respiratory diseases						
Bronchitis; not specified as acute or chronic	35075 (5.1)	11539 (6.4)	0.05	9616 (6.3)	9743 (6.4)	0.003
Simple and mucopurulent chronic bronchitis	2520 (0.4)	661 (0.4)	3.00E-04	596 (0.4)	618 (0.4)	0.002
Unspecified chronic bronchitis	3204 (0.5)	722 (0.4)	0.01	653 (0.4)	694 (0.5)	0.004
Emphysema	10526 (1.5)	1863 (1.0)	0.05	1800 (1.2)	1841 (1.2)	0.002
Other chronic obstructive pulmonary disease	32681 (4.8)	6283 (3.5)	0.07	5892 (3.9)	6086 (4.0)	0.007
Asthma	79654 (11.7)	27763 (15.5)	0.1	23345 (15.3)	22375 (14.6)	0.02
Bronchiectasis	3537 (0.5)	805 (0.4)	0.01	766 (0.5)	742 (0.5)	0.002
Nicotine dependence	55991 (8.2)	15263 (8.5)	0.01	14323 (9.4)	14980 (9.8)	0.01
Psychiatric comorbidities						
Substance misuse	77411 (11.4)	18886 (10.5)	0.03	17694 (11.6)	18553 (12.1)	0.02
Psychotic disorders	9359 (1.4)	1327 (0.7)	0.06	1277 (0.8)	1320 (0.9)	0.003
Mood disorders	108333 (15.9)	22157 (12.3)	0.1	21671 (14.2)	21735 (14.2)	0.001

Anxiety disorders	137335 (20.2)	29633 (16.5)	0.09	28825 (18.9)	28507 (18.7)	0.005
Heart disease						
Ischemic heart diseases	65745 (9.7)	8063 (4.5)	0.2	8797 (5.8)	8001 (5.2)	0.02
Other forms of heart disease	127337 (18.7)	18710 (10.4)	0.2	19407 (12.7)	17962 (11.8)	0.03
Chronic kidney diseases						
Chronic kidney disease (CKD)	46479 (6.8)	5468 (3.0)	0.2	6025 (3.9)	5368 (3.5)	0.02
Hypertensive chronic kidney disease	26436 (3.9)	3091 (1.7)	0.1	3410 (2.2)	3053 (2.0)	0.02
Chronic liver disease						
Alcoholic liver disease	2817 (0.4)	335 (0.2)	0.04	363 (0.2)	333 (0.2)	0.004
Hepatic failure; not elsewhere classified	3637 (0.5)	436 (0.2)	0.05	492 (0.3)	422 (0.3)	0.008
Chronic hepatitis; not elsewhere classified	730 (0.1)	138 (0.08)	0.01	122 (0.08)	136 (0.09)	0.003
Fibrosis and cirrhosis of liver	7012 (1.0)	874 (0.5)	0.06	963 (0.6)	858 (0.6)	0.009
Fatty (change of) liver; not elsewhere classified	25758 (3.8)	3575 (2.0)	0.1	3689 (2.4)	3542 (2.3)	0.006
Chronic passive congestion of liver	3819 (0.6)	715 (0.4)	0.02	708 (0.5)	692 (0.5)	0.002
Portal hypertension	2861 (0.4)	329 (0.2)	0.04	386 (0.3)	318 (0.2)	0.009
Other specified diseases of liver	15404 (2.3)	2482 (1.4)	0.07	2548 (1.7)	2430 (1.6)	0.006
Cerebral infarction	15303 (2.2)	1651 (0.9)	0.1	1930 (1.3)	1629 (1.1)	0.02
Dementia						
Vascular dementia	2713 (0.4)	178 (0.1)	0.06	264 (0.2)	177 (0.1)	0.02
Dementia in other diseases classified elsewhere	5460 (0.8)	334 (0.2)	0.09	499 (0.3)	332 (0.2)	0.02
Unspecified dementia	11657 (1.7)	737 (0.4)	0.1	1056 (0.7)	732 (0.5)	0.03
Alzheimer disease	4430 (0.7)	255 (0.1)	0.08	394 (0.3)	255 (0.2)	0.02
Frontotemporal dementia	216 (0.03)	10 (0.006)	0.02	26 (0.02)	10 (0.007)	0.01
Dementia with Lewy bodies	352 (0.05)	28 (0.02)	0.02	38 (0.03)	28 (0.02)	0.004
Neoplasms						
Neoplasms (any)	133697 (19.6)	23651 (13.2)	0.2	23549 (15.4)	22229 (14.6)	0.02
Haematological cancer	7827 (1.1)	1572 (0.9)	0.03	1584 (1.0)	1473 (1.0)	0.007
Organ transplant						
Renal Transplantation Procedures	1619 (0.2)	164 (0.09)	0.04	225 (0.1)	158 (0.1)	0.01
Liver Transplantation Procedures	330 (0.05)	42 (0.02)	0.01	51 (0.03)	36 (0.02)	0.006
Psoriasis	7957 (1.2)	1574 (0.9)	0.03	1462 (1.0)	1525 (1.0)	0.004
Rheumatoid arthritis						
Rheumatoid arthritis with rheumatoid factor	2878 (0.4)	561 (0.3)	0.02	545 (0.4)	554 (0.4)	0.001
Other rheumatoid arthritis	10360 (1.5)	1869 (1.0)	0.04	1804 (1.2)	1842 (1.2)	0.002
Systemic lupus erythematosus (SLE)	3899 (0.6)	751 (0.4)	0.02	699 (0.5)	738 (0.5)	0.004
Disorders involving the immune mechanism	15884 (2.3)	3354 (1.9)	0.03	3167 (2.1)	3082 (2.0)	0.004

eTable 2 – Baseline characteristics for the matched subgroups of patients aged ≤ 16 years-old

	COVID-19	Influenza	SMD
Number	43231	43231	-
DEMOGRAPHICS			
Age; mean (SD); y	7.1 (4.9)	6.9 (4.6)	0.05
Sex; n (%)			
Female	20726 (47.9)	20515 (47.5)	0.01
Male	22500 (52.0)	22713 (52.5)	0.01
Other	10 (0.02)	10 (0.02)	0
Race; n (%)			
White	25632 (59.3)	25460 (58.9)	0.008
Black or African American	7616 (17.6)	7509 (17.4)	0.007
Asian	1050 (2.4)	1003 (2.3)	0.007
American Indian or Alaska Native	165 (0.4)	160 (0.4)	0.002
Native Hawaiian or Other Pacific Islander	94 (0.2)	117 (0.3)	0.01
Unknown	8674 (20.1)	8982 (20.8)	0.02
Ethnicity; n (%)			
Hispanic or Latino	7567 (17.5)	7615 (17.6)	0.003
Not Hispanic of Latino	24954 (57.7)	24614 (56.9)	0.02
Unknown	10710 (24.8)	11002 (25.4)	0.02
Socioeconomic deprivation; n (%)	87 (0.2)	80 (0.2)	0.004
COMORBIDITIES; n (%)			
Overweight and obesity	1900 (4.4)	1947 (4.5)	0.005
Hypertensive disease	392 (0.9)	408 (0.9)	0.004
Diabetes mellitus			
Type 1 diabetes mellitus	177 (0.4)	181 (0.4)	0.001
Type 2 diabetes mellitus	141 (0.3)	149 (0.3)	0.003
Chronic lower respiratory diseases			
Bronchitis; not specified as acute or chronic	964 (2.2)	881 (2.0)	0.01
Simple and mucopurulent chronic bronchitis	29 (0.07)	26 (0.06)	0.003
Unspecified chronic bronchitis	20 (0.05)	13 (0.03)	0.008
Emphysema	10 (0.02)	10 (0.02)	0
Other chronic obstructive pulmonary disease	78 (0.2)	63 (0.1)	0.009
Asthma	5493 (12.7)	5302 (12.3)	0.01
Bronchiectasis	29 (0.07)	30 (0.07)	9.00E-04
Nicotine dependence	35 (0.08)	31 (0.07)	0.003
Psychiatric comorbidities			
Substance misuse	123 (0.3)	117 (0.3)	0.003
Psychotic disorders	23 (0.05)	30 (0.07)	0.007
Mood disorders	655 (1.5)	786 (1.8)	0.02
Anxiety disorders	2319 (5.4)	2391 (5.5)	0.007
Heart disease			
Ischemic heart diseases	57 (0.1)	56 (0.1)	6.00E-04

Other forms of heart disease	1186 (2.7)	1252 (2.9)	0.009
Chronic kidney diseases			
Chronic kidney disease (CKD)	110 (0.3)	110 (0.3)	0
Hypertensive chronic kidney disease	32 (0.07)	33 (0.08)	8.00E-04
Chronic liver disease			
Alcoholic liver disease	0 (0.0)	0 (0.0)	NA
Hepatic failure; not elsewhere classified	23 (0.05)	22 (0.05)	0.001
Chronic hepatitis; not elsewhere classified	10 (0.02)	10 (0.02)	0
Fibrosis and cirrhosis of liver	18 (0.04)	16 (0.04)	0.002
Fatty (change of) liver; not elsewhere classified	46 (0.1)	62 (0.1)	0.01
Chronic passive congestion of liver	27 (0.06)	26 (0.06)	9.00E-04
Portal hypertension	12 (0.03)	13 (0.03)	0.001
Other specified diseases of liver	67 (0.2)	62 (0.1)	0.003
Cerebral infarction	34 (0.08)	36 (0.08)	0.002
Dementia			
Vascular dementia	0 (0.0)	0 (0.0)	NA
Dementia in other diseases classified elsewhere	10 (0.02)	10 (0.02)	0
Unspecified dementia	10 (0.02)	10 (0.02)	0
Alzheimer disease	10 (0.02)	0 (0.0)	0.02
Frontotemporal dementia	0 (0.0)	0 (0.0)	NA
Dementia with Lewy bodies	0 (0.0)	0 (0.0)	NA
Neoplasms			
Neoplasms (any)	1659 (3.8)	1721 (4.0)	0.007
Haematological cancer	134 (0.3)	141 (0.3)	0.003
Organ transplant			
Renal Transplantation Procedures	10 (0.02)	11 (0.03)	0.001
Liver Transplantation Procedures	10 (0.02)	10 (0.02)	0
Psoriasis	65 (0.1)	68 (0.2)	0.002
Rheumatoid arthritis			
Rheumatoid arthritis with rheumatoid factor	10 (0.02)	10 (0.02)	0
Other rheumatoid arthritis	15 (0.04)	10 (0.02)	0.007
Systemic lupus erythematosus (SLE)	10 (0.02)	11 (0.03)	0.001
Disorders involving the immune mechanism	376 (0.9)	356 (0.8)	0.005

eTable 3 – Baseline characteristics for the matched subgroups of patients aged > 16 years-old

	COVID-19	**Influenza**	**SMD**
Number	108116	108116	-
DEMOGRAPHICS			
Age; mean (SD); y	42.6 (17.7)	42.5 (17.8)	0.004
Sex; n (%)			
Female	64086 (59.3)	64133 (59.3)	9.00E-04
Male	44008 (40.7)	43968 (40.7)	8.00E-04
Other	22 (0.02)	15 (0.01)	0.005
Race; n (%)			
White	73643 (68.1)	73803 (68.3)	0.003
Black or African American	18431 (17.0)	18110 (16.8)	0.008
Asian	3931 (3.6)	3736 (3.5)	0.01
American Indian or Alaska Native	457 (0.4)	474 (0.4)	0.002
Native Hawaiian or Other Pacific Islander	167 (0.2)	177 (0.2)	0.002
Unknown	11487 (10.6)	11816 (10.9)	0.01
Ethnicity; n (%)			
Hispanic or Latino	9310 (8.6)	9241 (8.5)	0.002
Not Hispanic of Latino	75694 (70.0)	76078 (70.4)	0.008
Unknown	23112 (21.4)	22797 (21.1)	0.007
Socioeconomic deprivation; n (%)	806 (0.7)	865 (0.8)	0.006
COMORBIDITIES; n (%)			
Overweight and obesity	21543 (19.9)	21868 (20.2)	0.008
Hypertensive disease	30459 (28.2)	30715 (28.4)	0.005
Diabetes mellitus			
Type 1 diabetes mellitus	1813 (1.7)	1954 (1.8)	0.01
Type 2 diabetes mellitus	12616 (11.7)	12849 (11.9)	0.007
Chronic lower respiratory diseases			
Bronchitis; not specified as acute or chronic	8903 (8.2)	9685 (9.0)	0.03
Simple and mucopurulent chronic bronchitis	596 (0.6)	672 (0.6)	0.009
Unspecified chronic bronchitis	656 (0.6)	717 (0.7)	0.007
Emphysema	1872 (1.7)	1916 (1.8)	0.003
Other chronic obstructive pulmonary disease	6016 (5.6)	6379 (5.9)	0.01
Asthma	16438 (15.2)	16958 (15.7)	0.01
Bronchiectasis	718 (0.7)	769 (0.7)	0.006
Nicotine dependence	14821 (13.7)	15530 (14.4)	0.02
Psychiatric comorbidities			
Substance misuse	18376 (17.0)	19158 (17.7)	0.02
Psychotic disorders	1237 (1.1)	1318 (1.2)	0.007
Mood disorders	21618 (20.0)	22043 (20.4)	0.01
Anxiety disorders	27211 (25.2)	27454 (25.4)	0.005
Heart disease			
Ischemic heart diseases	8173 (7.6)	8377 (7.7)	0.007

Other forms of heart disease	17146 (15.9)	17703 (16.4)	0.01
Chronic kidney diseases			
Chronic kidney disease (CKD)	5451 (5.0)	5627 (5.2)	0.007
Hypertensive chronic kidney disease	3102 (2.9)	3164 (2.9)	0.003
Chronic liver disease			
Alcoholic liver disease	347 (0.3)	346 (0.3)	2.00E-04
Hepatic failure; not elsewhere classified	405 (0.4)	419 (0.4)	0.002
Chronic hepatitis; not elsewhere classified	133 (0.1)	136 (0.1)	8.00E-04
Fibrosis and cirrhosis of liver	830 (0.8)	891 (0.8)	0.006
Fatty (change of) liver; not elsewhere classified	3526 (3.3)	3641 (3.4)	0.006
Chronic passive congestion of liver	655 (0.6)	688 (0.6)	0.004
Portal hypertension	323 (0.3)	325 (0.3)	3.00E-04
Other specified diseases of liver	2378 (2.2)	2485 (2.3)	0.007
Cerebral infarction	1594 (1.5)	1682 (1.6)	0.007
Dementia			
Vascular dementia	193 (0.2)	185 (0.2)	0.002
Dementia in other diseases classified elsewhere	400 (0.4)	346 (0.3)	0.009
Unspecified dementia	791 (0.7)	756 (0.7)	0.004
Alzheimer disease	302 (0.3)	265 (0.2)	0.007
Frontotemporal dementia	19 (0.02)	13 (0.01)	0.005
Dementia with Lewy bodies	24 (0.02)	29 (0.03)	0.003
Neoplasms			
Neoplasms (any)	21441 (19.8)	21916 (20.3)	0.01
Haematological cancer	1395 (1.3)	1464 (1.4)	0.006
Organ transplant			
Renal Transplantation Procedures	148 (0.1)	161 (0.1)	0.003
Liver Transplantation Procedures	45 (0.04)	38 (0.04)	0.003
Psoriasis	1432 (1.3)	1534 (1.4)	0.008
Rheumatoid arthritis			
Rheumatoid arthritis with rheumatoid factor	539 (0.5)	593 (0.5)	0.007
Other rheumatoid arthritis	1791 (1.7)	1913 (1.8)	0.009
Systemic lupus erythematosus (SLE)	751 (0.7)	760 (0.7)	0.001
Disorders involving the immune mechanism	2817 (2.6)	2913 (2.7)	0.006

eTable 4 – Baseline characteristics for the matched subgroups of patients not hospitalised at the time of the index infection

	COVID-19	Influenza	SMD
Number	139490	139490	-
DEMOGRAPHICS			
Age; mean (SD); y	30.1 (20.5)	29.1 (21.7)	0.05
Sex; n (%)			
Female	77270 (55.4)	78245 (56.1)	0.01
Male	62194 (44.6)	61231 (43.9)	0.01
Other	26 (0.02)	14 (0.01)	0.007
Race; n (%)			
White	90989 (65.2)	88887 (63.7)	0.03
Black or African American	22991 (16.5)	22809 (16.4)	0.004
Asian	4987 (3.6)	4714 (3.4)	0.01
American Indian or Alaska Native	570 (0.4)	588 (0.4)	0.002
Native Hawaiian or Other Pacific Islander	283 (0.2)	313 (0.2)	0.005
Unknown	19670 (14.1)	22179 (15.9)	0.05
Ethnicity; n (%)			
Hispanic or Latino	17968 (12.9)	19530 (14.0)	0.03
Not Hispanic of Latino	99433 (71.3)	95132 (68.2)	0.07
Unknown	22089 (15.8)	24828 (17.8)	0.05
Socioeconomic deprivation; n (%)	650 (0.5)	652 (0.5)	2.00E-04
COMORBIDITIES; n (%)			
Overweight and obesity	20069 (14.4)	19889 (14.3)	0.004
Hypertensive disease	25374 (18.2)	23326 (16.7)	0.04
Diabetes mellitus			
Type 1 diabetes mellitus	1587 (1.1)	1518 (1.1)	0.005
Type 2 diabetes mellitus	10416 (7.5)	9317 (6.7)	0.03
Chronic lower respiratory diseases			
Bronchitis; not specified as acute or chronic	8892 (6.4)	8688 (6.2)	0.006
Simple and mucopurulent chronic bronchitis	464 (0.3)	472 (0.3)	0.001
Unspecified chronic bronchitis	469 (0.3)	501 (0.4)	0.004
Emphysema	1099 (0.8)	1101 (0.8)	2.00E-04
Other chronic obstructive pulmonary disease	3572 (2.6)	3594 (2.6)	0.001
Asthma	20655 (14.8)	19735 (14.1)	0.02
Bronchiectasis	518 (0.4)	494 (0.4)	0.003
Nicotine dependence	11662 (8.4)	12390 (8.9)	0.02
Psychiatric comorbidities			
Substance misuse	14523 (10.4)	15365 (11.0)	0.02
Psychotic disorders	964 (0.7)	979 (0.7)	0.001
Mood disorders	18833 (13.5)	18930 (13.6)	0.002
Anxiety disorders	25406 (18.2)	25338 (18.2)	0.001
Heart disease			

Ischemic heart diseases	5780 (4.1)	5200 (3.7)	0.02
Other forms of heart disease	14644 (10.5)	13466 (9.7)	0.03
Chronic kidney diseases			
Chronic kidney disease (CKD)	3831 (2.7)	3429 (2.5)	0.02
Hypertensive chronic kidney disease	2003 (1.4)	1835 (1.3)	0.01
Chronic liver disease			
Alcoholic liver disease	228 (0.2)	208 (0.1)	0.004
Hepatic failure; not elsewhere classified	273 (0.2)	223 (0.2)	0.009
Chronic hepatitis; not elsewhere classified	102 (0.07)	98 (0.07)	0.001
Fibrosis and cirrhosis of liver	661 (0.5)	571 (0.4)	0.01
Fatty (change of) liver; not elsewhere classified	3230 (2.3)	3009 (2.2)	0.01
Chronic passive congestion of liver	522 (0.4)	524 (0.4)	2.00E-04
Portal hypertension	218 (0.2)	192 (0.1)	0.005
Other specified diseases of liver	2073 (1.5)	1983 (1.4)	0.005
Cerebral infarction	1261 (0.9)	1097 (0.8)	0.01
Dementia			
Vascular dementia	170 (0.1)	111 (0.08)	0.01
Dementia in other diseases classified elsewhere	293 (0.2)	192 (0.1)	0.02
Unspecified dementia	560 (0.4)	373 (0.3)	0.02
Alzheimer disease	220 (0.2)	141 (0.1)	0.02
Frontotemporal dementia	12 (0.009)	10 (0.007)	0.002
Dementia with Lewy bodies	27 (0.02)	17 (0.01)	0.006
Neoplasms			
Neoplasms (any)	20800 (14.9)	19294 (13.8)	0.03
Haematological cancer	1036 (0.7)	1024 (0.7)	0.001
Organ transplant			
Renal Transplantation Procedures	93 (0.07)	93 (0.07)	0
Liver Transplantation Procedures	26 (0.02)	20 (0.01)	0.003
Psoriasis	1359 (1.0)	1329 (1.0)	0.002
Rheumatoid arthritis			
Rheumatoid arthritis with rheumatoid factor	435 (0.3)	459 (0.3)	0.003
Other rheumatoid arthritis	1467 (1.1)	1471 (1.1)	3.00E-04
Systemic lupus erythematosus (SLE)	550 (0.4)	608 (0.4)	0.006
Disorders involving the immune mechanism	2396 (1.7)	2376 (1.7)	0.001

eTable 5 – Baseline characteristics for the matched subgroups of patients hospitalised at the time of the index infection

	COVID-19	Influenza	SMD
Number	11090	11090	-
DEMOGRAPHICS			
Age; mean (SD); y	46.1 (23.8)	46.3 (26.6)	0.01
Sex; n (%)			
Female	5928 (53.5)	5927 (53.4)	2.00E-04
Male	5161 (46.5)	5162 (46.5)	2.00E-04
Other	10 (0.09)	10 (0.09)	0
Race; n (%)			
White	7201 (64.9)	7307 (65.9)	0.02
Black or African American	2377 (21.4)	2310 (20.8)	0.01
Asian	318 (2.9)	312 (2.8)	0.003
American Indian or Alaska Native	36 (0.3)	55 (0.5)	0.03
Native Hawaiian or Other Pacific Islander	15 (0.1)	11 (0.1)	0.01
Unknown	1143 (10.3)	1095 (9.9)	0.01
Ethnicity; n (%)			
Hispanic or Latino	1052 (9.5)	985 (8.9)	0.02
Not Hispanic of Latino	8208 (74.0)	8226 (74.2)	0.004
Unknown	1830 (16.5)	1879 (16.9)	0.01
Socioeconomic deprivation; n (%)	335 (3.0)	305 (2.8)	0.02
COMORBIDITIES; n (%)			
Overweight and obesity	2879 (26.0)	2785 (25.1)	0.02
Hypertensive disease	5781 (52.1)	5631 (50.8)	0.03
Diabetes mellitus			
Type 1 diabetes mellitus	548 (4.9)	517 (4.7)	0.01
Type 2 diabetes mellitus	3016 (27.2)	2924 (26.4)	0.02
Chronic lower respiratory diseases			
Bronchitis; not specified as acute or chronic	952 (8.6)	962 (8.7)	0.003
Simple and mucopurulent chronic bronchitis	128 (1.2)	138 (1.2)	0.008
Unspecified chronic bronchitis	194 (1.7)	189 (1.7)	0.003
Emphysema	730 (6.6)	726 (6.5)	0.001
Other chronic obstructive pulmonary disease	2475 (22.3)	2504 (22.6)	0.006
Asthma	2453 (22.1)	2464 (22.2)	0.002
Bronchiectasis	267 (2.4)	258 (2.3)	0.005
Nicotine dependence	2604 (23.5)	2551 (23.0)	0.01
Psychiatric comorbidities			
Substance misuse	3213 (29.0)	3124 (28.2)	0.02
Psychotic disorders	314 (2.8)	323 (2.9)	0.005
Mood disorders	2814 (25.4)	2737 (24.7)	0.02
Anxiety disorders	3126 (28.2)	3055 (27.5)	0.01
Heart disease			

Ischemic heart diseases	2764 (24.9)	2724 (24.6)	0.008
Other forms of heart disease	4564 (41.2)	4470 (40.3)	0.02
Chronic kidney diseases			
Chronic kidney disease (CKD)	1923 (17.3)	1912 (17.2)	0.003
Hypertensive chronic kidney disease	1224 (11.0)	1195 (10.8)	0.008
Chronic liver disease			
Alcoholic liver disease	122 (1.1)	126 (1.1)	0.003
Hepatic failure; not elsewhere classified	228 (2.1)	205 (1.8)	0.01
Chronic hepatitis; not elsewhere classified	39 (0.4)	36 (0.3)	0.005
Fibrosis and cirrhosis of liver	310 (2.8)	291 (2.6)	0.01
Fatty (change of) liver; not elsewhere classified	551 (5.0)	525 (4.7)	0.01
Chronic passive congestion of liver	169 (1.5)	160 (1.4)	0.007
Portal hypertension	129 (1.2)	129 (1.2)	0
Other specified diseases of liver	405 (3.7)	429 (3.9)	0.01
Cerebral infarction	515 (4.6)	517 (4.7)	9.00E-04
Dementia			
Vascular dementia	78 (0.7)	63 (0.6)	0.02
Dementia in other diseases classified elsewhere	145 (1.3)	138 (1.2)	0.006
Unspecified dementia	367 (3.3)	353 (3.2)	0.007
Alzheimer disease	114 (1.0)	111 (1.0)	0.003
Frontotemporal dementia	10 (0.09)	10 (0.09)	0
Dementia with Lewy bodies	10 (0.09)	11 (0.1)	0.003
Neoplasms			
Neoplasms (any)	2853 (25.7)	2772 (25.0)	0.02
Haematological cancer	428 (3.9)	440 (4.0)	0.006
Organ transplant			
Renal Transplantation Procedures	66 (0.6)	67 (0.6)	0.001
Liver Transplantation Procedures	20 (0.2)	19 (0.2)	0.002
Psoriasis	196 (1.8)	176 (1.6)	0.01
Rheumatoid arthritis			
Rheumatoid arthritis with rheumatoid factor	95 (0.9)	95 (0.9)	0
Other rheumatoid arthritis	362 (3.3)	358 (3.2)	0.002
Systemic lupus erythematosus (SLE)	120 (1.1)	132 (1.2)	0.01
Disorders involving the immune mechanism	708 (6.4)	696 (6.3)	0.004

References

1 Taquet M, Dercon Q, Luciano S, Geddes JR, Husain M, Harrison PJ. Incidence, co-occurrence, and evolution of long-COVID features: A 6-month retrospective cohort study of 273,618 survivors of COVID-19. *PLoS Med* 2021; **18**: e1003773.

2 Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health* 2016; **37**: 61–81.

3 Cowie MR, Blomster JI, Curtis LH, *et al.* Electronic health records to facilitate clinical research. *Clin Res Cardiol* 2017; **106**: 1–9.

4 Jetley G, Zhang H. Electronic health records in IS research: Quality issues, essential thresholds and remedial actions. *Decis Support Syst* 2019; **126**: 113137.

5 de Lusignan S, Dorward J, Correa A, *et al.* Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study. *Lancet Infect Dis* 2020; published online May 15. DOI:10.1016/S1473-3099(20)30371-6.

6 Zhang J-J, Dong X, Cao Y-Y, *et al.* Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. *Allergy* 2020. https://onlinelibrary.wiley.com/doi/abs/10.1111/all.14238.

7 Chen N, Zhou M, Dong X, *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *Lancet* 2020; **395**: 507–13.

8 Wang QQ, Kaelber DC, Xu R, Volkow ND. COVID-19 risk and outcomes in patients with substance use disorders: analyses from electronic health records in the United States. *Mol Psychiatry* 2021; **26**: 30–9.

9 Williamson EJ, Walker AJ, Bhaskaran K, *et al.* OpenSAFELY: factors associated with COVID-19 death in 17 million patients. *Nature* 2020; published online July 8. DOI:10.1038/s41586-020-2521-4.

10 Taquet M, Luciano S, Geddes JR, Harrison PJ. Bidirectional associations between COVID-19 and psychiatric disorder: retrospective cohort studies of 62 354 COVID-19 cases in the USA. *Lancet Psychiatry* 2021; **8**: 130–40.

11 Taquet M, Geddes JR, Husain M, Luciano S, Harrison PJ. 6-month neurological and psychiatric outcomes in 236 379 survivors of COVID-19: a retrospective cohort study using electronic health records. *Lancet Psychiatry* 2021; **8**: 416–27.

12 Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 1994; **81**: 515.

13 Royston P, Parmar MKB. Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine. 2002; **21**: 2175–97.

14 Liu X-R, Pawitan Y, Clements M. Parametric and penalized generalized survival models. *Stat Methods Med Res* 2018; **27**: 1531–46.