nature portfolio

Corresponding author(s): Fei Wang

Last updated by author(s): Mar 15, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No particular software was used for collecting the data.
Data analysis	For reproducibility, our codes are available at https://github.com/calvin-zcx/pasc_phenotype. See https://zenodo.org/account/settings/github/repository/calvin-zcx/pasc_phenoty for citing the code. We used Python 3.9, python package lifelines-0.2666 , scikit-learn-0.2318 and the Clinical Classifications Software Refined (CCSR) v2022-1, the Elixhauser Comorbidity Indices Refined for ICD-10-CM.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The INSIGHT data can be requested through https://insightcrn.org/. The OneFlorida+ data can be requested through https://onefloridaconsortium.org. Both the INSIGHT and the OneFlorida+ data are HIPAA-limited. Therefore, data use agreements must be established with the INSIGHT and OneFlorida+ networks.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	This is a retrospective secondary analysis of de-identified patient records from two large electronic health record (EHR) cohorts. Summary statistics on sex distributions within different cohorts were reported in Table 1 and Supplementary Data 3. Summary statistics on PASC distributions on sex were reported in Figure 4, extended Fig 2,4,5.
Population characteristics	We summarized the baseline characteristics of both the INSIGHT cohort and OneFlorida+ cohort in Table 1 from information that was available on patients in clinical data; demographic information was collected from patients when they registered for care within the healthcare systems. We observed significant differences between the two cohorts regarding age, gender, race, area deprivation index, and outbreak waves. The INSIGHT cohort contained SARS-CoV-2 infected patients mainly from the New York metropolitan area with the median area deprivation index (ADI, rankings from 1 to 100, with 1 and 100 indicating the lowest and highest level of disadvantage)17 15 (6-24) in the SARS-CoV-2 infected patient group, indicating fewer disadvantaged neighborhoods than the OneFlorida+ cohort whose median ADI was 58 (41-76). Indeed, the OneFlorida + cohort consisted of a mixture of urban, suburban, and rural populations in Florida and selected cities in Georgia and Alabama (see Methods). The median age of SARS-CoV-2 infected patients in the INSIGHT cohort was 55 (38-68), older than the OneFlorida+ cohort with a median age of 50 (34-64). Plus, more female SARS-CoV-2 infected patients were in the OneFlorida+ cohort (62.7%) than in the INSIGHT cohort (58.6%). The INSIGHT cohort also had a more diverse population with 34.7% white and 54.9% others (Asian and others including American Indian or Alaska Native, Native Hawaiian or other Pacific Islander, multiple races, etc.); the OneFlorida+ cohort had a majority of patients identifying as White race (51.0%). Additionally, there is a higher proportion of patients infected early in the pandemic in the INSIGHT cohort (31.8% of all infected patients were from March 2020 to June 2020) compared to the OneFlorida+ cohort (9.1% of cases were from March 2020 to June 2020). Different temporal patterns of new cases per month across two cohorts are illustrated in Extended Data Fig. 1. The two networks also differed in care settings connected to patient encounters and treatments utilized for infected patients (e.g., more inpatient visits and more prescriptions of corticosteroids in the OneFlorida+ cohort than in the INSIGHT cohort).
Recruitment	This is a retrospective secondary analysis of EHR data and no patient recruitment activities are involved.
Ethics oversight	The use of the INSIGHT data was approved by the Institutional Review Board (IRB) of Weill Cornell Medicine following protocol 21-10-95-380 with title "Adult PCORnet-PASC Response to the Proposed Revised Milestones for the PASC EHR/ ORWD Teams (RECOVER)". The use of the OneFlorida+ data for this study was approved under the University of Florida IRB number IRB202001831.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our study is based two large scale patient-centered clinical research networks (PCORnet), INSIGHT and OneFlorida+. From 5,346,357 INSIGHT patients and 19,207,300 OneFlorida+ patients, we included 1,402,348 adult patients from the INSIGHT clinical research network (CRN) and 718,199 adult patients from the OneFlorida+ CRN who tested positive for SARS-CoV-2 on viral tests between March 01, 2020 and November 30, 2021. After applying eligibility criteria, the INSIGHT cohort finally contained 35,275 adult patients (age ≥ 20) with lab-confirmed SARS-CoV-2 infection who survived the first 30 days of infection from March 2020 to November 2021 in NYC and 326,126 eligible non-infected controls. The OneFlorida+ finally had 22,341 eligible lab-confirmed SARS-CoV-2 positive patients who survived the first 30 days of infection during the same period in Florida, Georgia, and Alabama and 177,010 non-infected controls.
Data exclusions	For the both cohorts, adult patients (age ≥ 20) with at least one SARS-CoV-2 polymerase-chain-reaction (PCR) or antigen laboratory test (Supplemental Table 1) between March 01, 2020 and November 30, 2021 were selected. Then we chose the patients who had at least one positive test and had at least one potential PASC conditions in the follow-up (or post-acute infection) period defined as below. We further made sure those potential PASC conditions were new incidences in the follow-up period by excluding patients who had any of them in both baseline and follow-up periods. The overall inclusion-exclusion cascade was shown in Figure 1, and the relevant definitions are provided below. Index date: the date of the first COVID-19 positive test. Baseline period: from 3 years to one week prior to the index date. Follow-up (post-acute infection) period: from 31 days after the index date to the day of documented death, last record in the database, 180 days after baseline, or the end of our observational window (Nov. 30, 2021), whichever came first.

Replication	The main analysis was conducted on the INSIGHT cohort and a replication analysis was done on the OneFlorida+ cohort.
Randomization	This is a retrospective analysis based on clustering and no randomization procedure was involved as no treatment effect was assessed. For the contrast analysis with COVID-19 patients, a similarity based matching procedure was performed for identifying appropriate COVID-19 patients to make fair comparisons. The matching covariates include the following • Demographics: age (20-<40 years, 40-<55 years, 55-<65 years, 65-<75 years, 75-<85 years, 85+ years),gender (female, male, other/missing), race (Asian, Black or African American, White, other, missing), ethnicity (Hispanic, not Hispanic, other/missing) which were categorized into different groups. • The area deprivation index (10-rank bins of national ADI) for capturing socioeconomic disadvantage of patients' neighborhood13. • Index date for considering the effect of different stages of pandemic, which was binned into different time intervals (March 2020 – June 2020, July 2020 – October 2020, November 2020 - February 2021, March 2021 – June 2021, July 2021 – November 2021). • Medical utilizations measured by numbers of inpatient, outpatient, and emergency encounters in the baseline period (binned into 0 visit, 1 or 2 visits, 3 or 4 visits, 5+ visits for each encounter type). • The body mass index (BMI) was categorized into underweight (<18.5 kg/m2), normal weight (18.5 kg/m2 – 24.9 kg/m2), overweight (25.0 kg/m2– 29.9 kg/m2), obesity (>= 30.0 kg/m2), and missing according to the CDC guideline for adults30 • Coexisting conditions including comorbidities and medications based on a tailored list of the Elixhauser comorbidities. We defined the patient having a particular condition if he/she had at least two related records during the baseline period. We built a propensity score (PS) model—the probability of assignment of a particular exposure group conditioned on baseline covariates—for each target outcome. Based on the estimated PS values, we then used stabilized inverse propensity score weighting (IPTW) 31 to re-weight patients in exposure and control groups, aiming to balance the two groups on baseline covariates after re-weighting. We further trimmed extreme weights beyond their 1st or 99th percentiles to control for potentially large weights to reduce variability. We used standardized mean difference (SMD) to quantify the goodness-of-balance of covariates over two groups The detailed description of this procedure was provided in Methods.
Blinding	This is a retrospective analysis and there is no blinding to the SARS-CoV-2 infection status.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging