Temporal-informed phenotyping scans the medical phenome to identify new diagnoses after recovery from COVID-19

By

Vern Eric Kerchberger, M.D.

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE
in
Biomedical Informatics
December 17, 2022

Nashville, Tennessee

Approved:

Wei-Qi Wei, M.D., Ph.D.

Lorraine B. Ware, M.D.

Colin Walsh, M.D., M.A.

QiPing Feng, Ph.D.

To my wife, Neha, and to my sons, Varen and Parks. You make everything worthwhile.

ACKNOWLEDGMENTS

I wish to thank the leadership of the Vanderbilt Biomedical Informatics Training Program including Cindy Gadd, Brad Malin, Rischelle Jenkins, and Kevin Johnson for providing me these educational opportunities and allowing me to work with the amazing faculty and students in the Vanderbilt DBMI. I also thank Josh Peterson for his support and guidance on developing this work into a published peer-reviewed manuscript. This work was made possible by funding from the National Institutes of Health (NIH T15 LM007450 and NIH K01 HL157755). I am eternally grateful to Lorraine Ware and Julie Bastarache for their inspiring mentorship, for encouraging me to pursue this program, and for fostering my career in academic medicine. I would not be here without their support and I look forward to a long career working together as colleagues. I also want to thank my committee for their mentorship through this process. In particular, I wish to thank my thesis mentor Wei-Qi Wei for his guidance, patience, and constant support of my research. I have come so far from when we first met in your office 4 years ago, thank you for believing in me.

To my parents, Vern Kerchberger and Mary Kane, for supporting my dreams throughout my life. Finally, I thank my wife Neha for her tireless support, love, and friendship. I couldn’t imagine having as much fun with anyone else.

TABLE OF CONTENTS

Page

Chapter

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION AND BACKGROUND

In the United States, passage of the Health Information Technology for Economic and Clinical Health (HITECH) Act in 2009 has led to a rapid uptake of electronic health records (EHRs) by hospitals and healthcare providers.[1,2] Whereas in 2008 fewer than 15% of US hospitals had EHRs, in 2019 nearly all non-federal US hospitals (96%) and most office-based physician practices (72%) reported use of EHRs.[3] This EHR expansion has resulted in massive increases in both the availability and density of longitudinal electronic health information for millions of patients, enabling novel investigations in multiple biomedical domains including epidemiology, outcomes research, clinical trial design, quality improvement, drug re-purposing, and precision medicine.[4–10]

High-throughput Phenotyping in the EHR with PheWAS

The volume and complexity of data accumulating in modern EHRs has necessitated development of novel strategies to translate this EHR data into computable phenotypes that have sufficient relevance and quality for biomedical research.[5,11,12] Multiple approaches have been employed to organize large volumes of health data both from structured formats including billing codes, vital signs, clinical laboratory test results, or medication prescriptions; and semi-structured or unstructured formats such as clinical reports, text from EHR notes, or images from diagnostic studies.[5,11–18] Among the many techniques described in the literature, phenome-wide association study (PheWAS) represents the prototypical high-throughput informatics analysis approach.[11,19–21] PheWAS was originally designed as a hypothesis-free method using EHR data to test the association between genetic variation at single-nucleotide polymorphisms (SNPs) and a broad set of clinical phenotypes represented as “phecodes”.[11,12,19,22] Phecodes are manually curated groupings of International Classification of Diseases (ICD) diagnosis codes, which are ubiquitous in modern EHR systems and capture a broad representation of all human medical conditions.[11,12,22] Each phecode groups one or more ICD codes for symptoms, findings, and/or diagnoses into a clinically meaningful phenotype, and also encodes a control definition to specify ICD codes that should not be present among “controls” (Table 1). Although the original phecode system used ICD codes from the Ninth Edition, Clinical Modification (ICD-9-CM), a mapping of ICD-10-CM to phecodes also exists and has demonstrated similar performance to the phecode mapping using ICD-9-CM.[22] The power of PheWAS was initially demonstrated in a seminal 2010 study by Denny et. al. that replicated four of seven known SNP-disease associations and identified 19 novel associations using a cohort of 6,005 genotyped European-American patients.[19] Subsequent studies have confirmed the capability of PheWAS to both replicate known SNP-phenotype associations and discover new candidate associations,[20,21,23] and established the superiority of phecodes over alternative high-throughput phenotyping approaches.[24] The PheWAS methodology has also been extended to study associations between EHR-derived phenotypes with non-genetic variables including clinical laboratory tests,[25,26] healthcare costs,[27] sleep habits,[28] chronic disease severity,[29] racial disparities,[30] and occupation.[31]

Table 1. Comparison of diagnosis coding schemas for high-throughput phenotyping

Code set	ICD-9-CM / ICD-10 CM	Phecodes / PheWAS
Number of codes / phenotypes	>15,000 (ICD-9-CM) >80,000 (ICD-10-CM)	1867
Embedded controls	No	Yes
Control definition	All patients except those with relevant “top” code	All patients except those with exclusion code(s)
Sex-specific exclusions	No	Yes
Example exclusion for “atrial fibrillation” ICD-9-CM 427.31 Phecode 427.12	ICD9CM: 427.* ICD10CM: I48.*	426-427.99

Adapted from Wei et. al.[24]

PheWAS using a phecode-based system has several advantages over alternative high-throughput phenotyping approaches (Table 1).[5,11,12,23,24] Firstly, as noted above the underlying source data (ICD codes) are ubiquitous in modern EHR systems and cover a large spectrum of known human diseases, affording greater portability of the approach across institutions.[11,12] Phecodes were designed by clinician-investigators with domain expertise in both medical taxonomy and adult internal medicine, which yielded a system of medically cognizable phenotypes built even from groups of disparate ICD codes. For example, the phecode 010 (Tuberculosis) contains codes from separate ICD chapters for both primary tuberculosis (ICD-9-CM 010-018; ICD-10-CM A15-A18) and for late effects of tuberculosis (ICD-9-CM 137; ICD-10-CM B90).[11] Phecodes also have a hierarchical structure which permits tailoring granularity level for a specific research question or for data availability, allowing PheWAS to examine both top codes like Diabetes Mellitus (phecode 250), and more specific leaf codes including Type I Diabetes Mellitus (phecode 250.1), Type II Diabetes Mellitus (phecode 250.2), Diabetic Retinopathy (phecode 250.7), etc.[11,12] Phecodes also have explicit control definitions which limits contamination of the control population by patients with potentially related conditions. Thus, in an EHR cohort study of atrial fibrillation (Phecode 427.12), “cases” would include all patients with a specific number of ICD codes for atrial fibrillation, while “controls” would be all patients with zero atrial fibrillation codes who also do not have a code for potentially related diagnoses such as atrial flutter (phecode 427.22), palpitations (phecode 427.9), or having a cardiac pacemaker in situ (phecode 427.91).[5,24] Finally, PheWAS has a freely-available *R* software package which facilitates dissemination across institutions and databases.[32]

The COVID-19 Pandemic as a Long-term Risk for Human Health

The coronavirus disease 2019 (COVID-19) pandemic is now in its third year. As of September 2022, there have been over 90 million confirmed cases in the United States and over 600 million cases worldwide, with many more cases likely going unreported.[33–35] Following the initial waves of widespread infection in Europe, Asia, and the United States, clinicians and researchers began reporting cohorts of COVID-19 survivors with both prolonged symptoms after initial illness as well as cohorts of survivors with new medical problems arising weeks or months after they ostensibly recovered.[36–52] Although no formal definition for this “post-COVID” syndrome (or alternatively termed “long COVID”) currently exists, multiple health authorities including the World Health Organization (WHO), British National Institute for Health Care and Excellent (NICE), and the US Centers for Disease Control and Prevention (CDC) have recognized that many COVID-19 survivors experience prolonged or new symptoms occurring at least four weeks following initial infection, that are demonstrably not due to active viral infection and infectivity.[53–55] Our understanding of the post-COVID syndrome remains incomplete, but it is likely common: a nation-wide study from the CDC reported up to 1 in 5 of adult COVID-19 survivors develop incident medical conditions potentially attributable to COVID-19.[52] The post-COVID syndrome is associated with a broad range of conditions and symptom clusters that can include physical symptoms, cognitive changes, or psychological ailments, along with a variety of cardiovascular, respiratory, renal, hematologic, endocrine, hepatic, gastrointestinal, mental health, and neurological diagnoses.[36–58] Although likely more frequent among survivors of severe disease, symptoms and new diagnoses attributable to the post-COVID syndrome are also common among suvivors of mild or even minimally symptomatic disease.[59–63] Duration of the post-COVID syndrome also appears variable, although some patients infected early in the pandemic still report lingering symptoms and lower overall health status as long as 2 years after their initial illness.[58]

Given the unprecedented scale of the COVID-19 pandemic, the protean nature of the post-COVID syndrome, and the millions of survivors subsequently interacting with modern healthcare systems, reliable high-throughput informatics methods like PheWAS could assist clinicians, researchers, and policymakers leverage the massive amounts of health data accumulating in EHR systems to improve our understanding of how COVID-19 infection will impact the long-term health of survivors. In particular, identifying new medical problems arising after recovery from acute COVID-19 would have many potential uses such as informing design of screening programs for survivors, assisting with modeling future costs to healthcare systems, and providing anticipatory guidance for patients and caregivers navigating the post-recovery phase of the disease. The extant studies applying PheWAS in the context of COVID-19 have focused on examining the comorbidity burdens associated with more severe COVID-19. Oetjens et al., applied PheWAS to EHR data from 1,604 COVID-19 patients receiving care at Geisinger Health, reporting high odds of pre-COVID chronic kidney disease, congestive hearth failure, diabetes, and chronic lung disease phenotypes among patients hospitalized with severe COVID-19.[64] Similarly, Salvatore et. al. used PheWAS on EHR data from 2,5852 COVID-19 positive adults receiving care a the University of Michigan to examine the association between pre-COVID comorbidity phenotypes with clinical COVID-19 outcomes including hospitalization, ICU admission, and death.[65] They similarly identified increased odds for multiple pre-morbid chronic disease phenotypes among COVID-19 patients with more severe outcomes.

As PheWAS was original designed to examine the effect of genetic variation (which aside from epigenetic modification rarely changes over time) on human health, most reported PheWAS studies have considered a patient’s medical phenome as a single longitudinal block or assessed a subset of ICD codes occurring around a specific event, such as a hospitalization or abnormal lab test, but without a focus on changes in patients’ phenotypes over the course of their medical history.[11,12,19–31] Currently, **no methods exist within the native PheWAS architecture to directly account for changes in patients’ medical conditions over time**.[32,66] Thus, new informatics methods are needed that can efficiently extract and identify new medical phenotypes arising after an acute temporal event within large-scale EHR data.

Motivation and Research Aims

The study presented in this thesis was motivated by the unmet need for an efficient high-throughput informatics method to ascertain how patients' health status change following an acute medical event, with COVID-19 infection serving as the demonstrative use case. This thesis is presented in three chapters. The first chapter describes the background and motivation of this research. The second chapter presents the development of a temporal-informed phenotyping approach and its application to identifying new medical phenotypes arising among COVID-19 survivors. The final chapter summarizes my experiments and discusses the limitations and future directions of this work.

CHAPTER 2

SCANNING THE MEDICAL PHENOME TO IDENTIFY NEW DIAGNOSES AFTER RECOVERY FROM COVID-19 IN A US COHORT

This manuscript was published in *Journal of the American Medical Informatics Association* as follows:

Kerchberger VE, Peterson JF, Wei WQ. Scanning the medical phenome to identify new diagnoses after recovery from COVID-19 in a US cohort. *J Am Med Inform Assoc*. 2022 Aug 25;ocac159. PMID: 36005898. PMCID: PMC9452157

doi: 10.1093/jamia/ocac159

Introduction

The coronavirus disease 2019 (COVID-19) pandemic continues to evolve, with more than 600 million confirmed cases worldwide over numerous waves.[33] Although most COVID-19 patients ultimately recover, many survivors report new medical problems arising after recovery from their acute illness.[38–51] With millions potentially at risk for long-term adverse health effects, methods to efficiently identify new medical problems occurring in survivors of COVID-19 or other acute medical events could be valuable for clinicians, researchers, and policymakers to improve identification of at-risk patients, discover new disease patterns, anticipate long-term consequences of acute illness on health systems, and plan for future pandemics.

Several database studies of medical conditions arising among COVID-19 survivors have been reported,[41,45,47,51] however these studies relied upon proprietary commercial claims or administrative data,[45] unique national databases,[41,47] or employed complex feature engineering and advanced statistical methods,[47,51] which potentially limits replication of research across institutions. Phenome-wide association study (PheWAS) is a high-throughput informatics framework initially developed to examine the effects of genetic variation on a wide range of physiological and clinical outcomes using electronic health records (EHR).[11,19,21,23,24] PheWAS has a well-documented R package incorporating feature engineering and analysis methods to facilitate study design and harmonization of research.[21,23,32] There also is increasing use of PheWAS to investigate the phenotypic consequences of non-genetic variables such as race, healthcare costs, or comorbidity burden.[12,25,27–30,64,65] While these characteristics appear favorable for enabling reproducible high-throughput studies of COVID-19 survivorship, the PheWAS feature engineering software does not account for temporal changes in a patient's medical conditions over time. To our knowledge prior PheWAS studies have not evaluated the development of new diagnoses after an acute medical event in real-world data.

In this study, we developed a temporal-informed phenotyping framework within the native PheWAS architecture to identify new diagnoses in the EHR occurring after an acute temporal event. Using this approach, we then systematically screened a large regional US registry to identify new medical conditions arising after recovery from acute COVID-19, hypothesizing that COVID-19 survivors have increased risk for new diagnoses ranging across the medical phenome.

Materials and Methods

Patient population and data sources

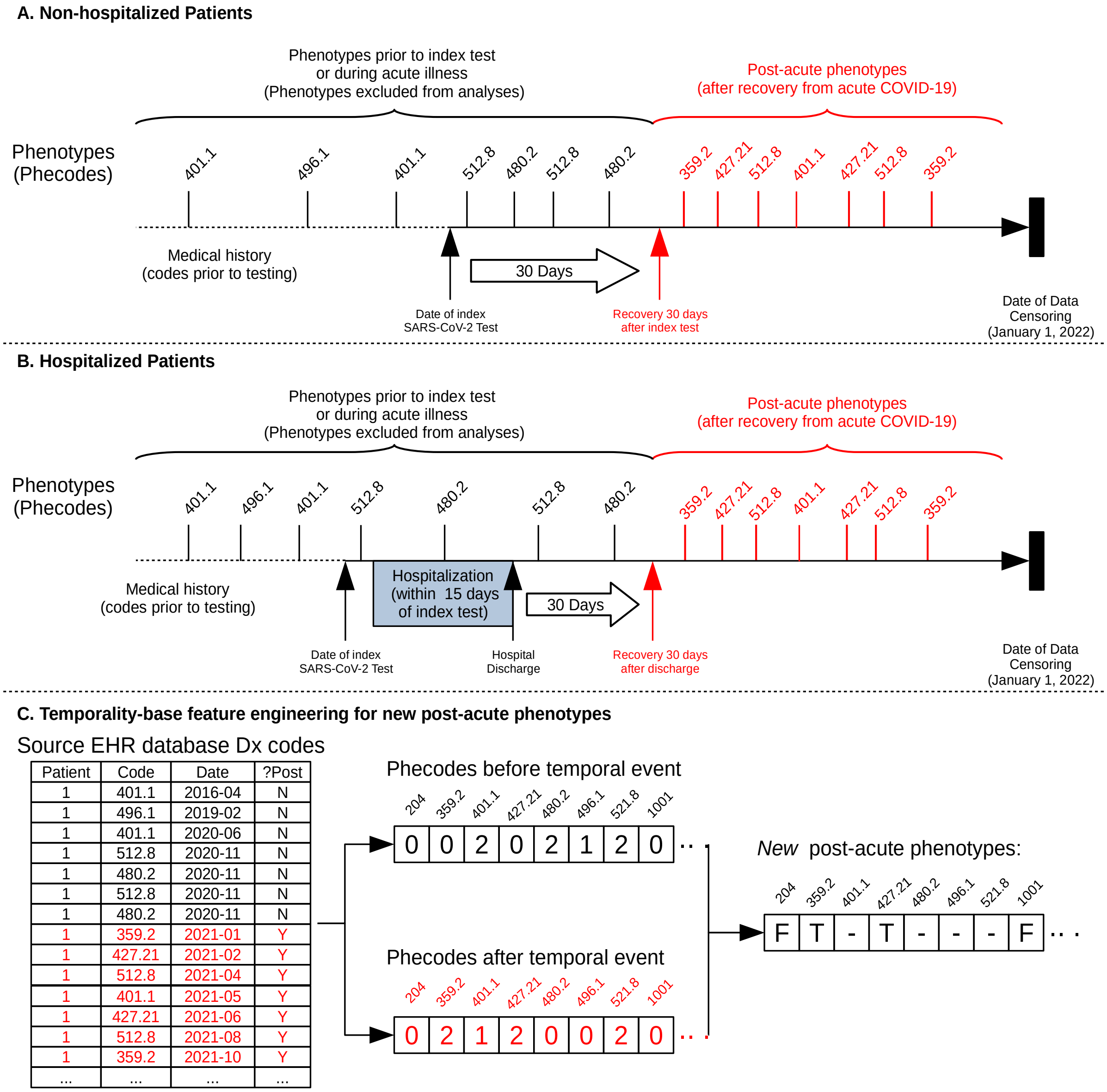
We used patient data from Vanderbilt University Medical Center's (VUMC) longitudinal COVID-19 EHR registry, and included all adults aged ≥18 years who had reverse transcription polymerase chain reaction (RT-PCR) testing for SARS-CoV-2 at VUMC from March 5, 2020 to November 1, 2021.[4,67] We excluded patients who had an ICD-10-CM code for laboratory-confirmed COVID-19 (U07.1) but never had a positive RT-PCR test at our institution, and patients who died before recovery from illness (defined below). Additional details on VUMC's COVID-19 registry database along with data cleaning methods are provided in Appendix A.

Defining post-acute COVID-19 in the EHR

Our temporal point of interest for identifying new medical problems was recovery from acute COVID-19. Using a generally-accepted definition for post-acute COVID-19 as four weeks after onset of symptoms,[38,39,47] we defined recovery from acute disease and transition to the post-acute phase as either 30 days after SARS-CoV-2 testing for non-hospitalized patients or 30 days after discharge for hospitalized patients (Figure 1). We used date of discharge for hospitalized patients as many critically ill COVID-19 patients have long hospital courses lasting weeks or months. We used the same definitions of the post-acute phase for never-infected patients to maintain congruent timing between the infected and uninfected groups.

Data collection

We collected ICD-9-CM and ICD-10-CM diagnosis codes entered into the EHR and grouped them into unique clinical phenotypes (phecodes) as commonly defined for PheWAS analyses.[22–24] We also collected vital sign values and results of common clinical laboratory tests obtained both prior to SARS-CoV-2 testing and after the post-acute phase. We censored data collection at January 1, 2022 so that the last patients tested in November 1, 2021 had at least 30 days of follow-up in the post-acute period. In keeping with usual practice for PheWAS, we defined "phenotype cases" as patients with a corresponding phecode on at least two separate days, and "phenotype controls" as patients with zero codes.[23,32] The native PheWAS feature engineering algorithm was used to automatically generate diagnosis-specific exclusion criteria for each phecode to mitigate contamination of the control group with potential cases. As an example: for an analysis of atrial fibrillation (phecode 427.21), patients who lack an atrial fibrillation diagnosis code but have potentially related diagnoses, signs, or symptoms of heart-rhythm disorders such as atrial flutter (phecode 427.22), palpitations (phecode 427.9), or cardiac pacemaker in situ (phecode 427.91) are excluded from the analysis rather than considered "phenotype controls".[12,22]


A. Non-hospitalized Patients
Phenotypes prior to index test or during acute illness (Phenotypes excluded from analyses)
Post-acute phenotypes (after recovery from acute COVID-19)
Phenotypes (Phecodes)
401.1
496.1
401.1
512.8
480.2
512.8
480.2
359.2
427.21
512.8
401.1
427.21
512.8
359.2
Medical history (codes prior to testing)
30 Days
Date of index SARS-CoV-2 Test
Recovery 30 days after index test
Date of Data Censoring (January 1, 2022)
B. Hospitalized Patients
Phenotypes prior to index test or during acute illness (Phenotypes excluded from analyses)
Post-acute phenotypes (after recovery from acute COVID-19)
Phenotypes (Phecodes)
401.1
496.1
401.1
512.8
480.2
512.8
480.2
359.2
427.21
512.8
401.1
427.21
512.8
359.2
Medical history (codes prior to testing)
Hospitalization (within 15 days of index test)
30 Days
Date of index SARS-CoV-2 Test
Hospital Discharge
Recovery 30 days after discharge
Date of Data Censoring (January 1, 2022)
C. Temporality-base feature engineering for new post-acute phenotypes
Source EHR database Dx codes
Phecodes before temporal event
204
359.2
401.1
427.21
480.2
496.1
521.8
1001
0 0 2 0 2 1 2 0 ··
Phecodes after temporal event
204
359.2
401.1
427.21
480.2
496.1
521.8
1001
0 2 1 2 0 0 2 0 ··
New post-acute phenotypes:
204
359.2
401.1
427.21
480.2
496.1
521.8
1001
F T - T - - - F ···

Patient	Code	Date	?Post
1	401.1	2016-04	N
1	496.1	2019-02	N
1	401.1	2020-06	N
1	512.8	2020-11	N
1	480.2	2020-11	N
1	512.8	2020-11	N
1	480.2	2020-11	N
1	359.2	2021-01	Y
1	427.21	2021-02	Y
1	512.8	2021-04	Y
1	401.1	2021-05	Y
1	427.21	2021-06	Y
1	512.8	2021-08	Y
1	359.2	2021-10	Y
...	...	...	...

Figure 1. Graphical timeline of data collection from electronic health record and phenotype encoding schematic.

Graphical timeline of index SARS-CoV-2 test, recovery, and phenotype case and control definitions for **A.** patients who were not hospitalized or **B.** were hospitalized around time of index SARS-CoV-2 test. Index date was defined as date of either first positive SARS-CoV-2 polymerase chain reaction (PCR) or first negative test for never-infected patients. Recovery date was defined as either **A.** 30 days after the index SARS-CoV-2 test in non-hospitalized patients or **B.** 30 days after hospital discharge in hospitalized patients. **C.** Schematic of temporal-informed phenotype feature engineering. The source EHR database was queried for diagnostic billing codes and the dataset was separated based on occurrence of codes before or after the temporal event (recovery date). Phecode feature engineering was applied to both "pre-event" and "post-event" datasets separately, then recombined to generate the final temporal-informed phenotypes. In this illustration, the patient is a temporal-informed case for phenotypes 359.2 and 427.21 (denoted as "T") as they had the corresponding diagnosis codes entered into the medical record on at least two separate dates after the temporal event, and did not have the diagnosis codes on a visit either before SARS-CoV-2 testing or during the acute phase. The patient is excluded from analyses of phenotypes 401.1, 480.2, 496.1, and 521.8 (denoted as "-") as they had those phecodes prior to the recovery date. The patient is a control for all phenotypes where they had zero codes in both the pre- and post-event datasets (e.g 204, 1001, and others; denoted as "F"). If the patient had a diagnosis-specific exclusion for a phecode in either dataset, the patient was excluded for that phecode in the temporal-informed phenotypes

Temporal-informed phenotype feature engineering

In assessing medical conditions arising after a temporal event, a naive phenotyping approach would be to use all diagnosis codes occurring after the event of interest. However, many medical diagnoses are chronic conditions for which patients receive repeated care. The naive phenotyping approach may not adequately distinguish new diagnoses from ongoing care for chronic diagnoses. To address this misclassification problem, we developed a temporal-informed phenotyping approach which separates each patient's medical phenome into two datasets based on occurrence of the diagnosis code relative to the event of interest (in this study, transition to the post-acute phase, Figure 1). We applied the PheWAS feature engineering method to the pre-event and post-event diagnosis code sets separately, and then recombined them using boolean logic to generate the temporal-informed phenotypes. In the final phenotype set, cases were patients with the phecode in post-event data and absent in pre-event data, while controls were patients where the phecode was absent in both sets. Patients who had an exclusion in either dataset or were a case in the pre-event data were converted to exclusions in the final temporal-informed phenotype dataset (Table 2).

Table 2. Boolean logic to generate temporal-informed phenotypes

Pre-event phenotype status	Post-event phenotype status	Temporal-informed phenotype	Comment
Control	Case	Case	Phenotype cases: diagnosis is new in post-event dataset.
Control	Control	Control	Phenotype controls: patients without diagnosis code in both pre-event and post-event datasets.
Case	Any	Exclude	Exclude patients with phenotype present prior to temporal event.
Exclude	Any	Exclude	Exclude patients with phenotype exclusion prior to temporal event.
Any	Exclude	Exclude	Exclude patients with phenotype exclusion after temporal event.

Statistical analyses and phenome-wide association testing

To assess the effects of our temporal-informed phenotyping on classifying PheWAS phenotypes, we compared case and control counts under the temporal-informed phenotyping approach to case and control counts under the naive approach. For each phecode, we calculated the case and control retention proportion $p_{\text{retention}}$ as shown in Equation 1:

$$p_{\text{retention}} = \frac{n_{\text{temporal-informed}}}{n_{\text{naive}}} \quad [1]$$

Where $n_{\text{temporal-informed}}$ is the phenotype case or control counts using temporal-informed phenotyping and n_{naive} is the phenotype case or control count under the naive approach. We compared case retention and control retention among phecode chapters (18 separate organ systems or categories based on ICD-9 chapters) using the nonparametric Mann-Whitney-U test. Tests of individual proportions were performed using the Chi-squared test.

In our analyses of temporal-informed phenotypes, the exposures of interest were (1) COVID-19 survivorship among all patients in the cohort, and (2) survivorship of severe COVID-19 (defined as admission to the hospital requiring supplemental oxygen) among SARS-CoV-2 positive patients.[68–70] We performed PheWAS using logistic regression to model the log-odds of developing each temporal-informed phenotype in the post-acute period given the presence or absence of the exposure of interest, adjusting for demographic and comorbidity covariates as Equation 2:

$$logit\, p(\mathrm{Y}_i = 1 | Exposure, Covariates) = \beta_0 + \beta_{EXPOSURE} \times Exposure + \beta_{COVAR} \times Covariates \quad [2]$$

where i = {1, …, n} phecodes with at least 10 phenotype cases in the cohort.[22,24] For vital signs and clinical laboratory tests, we modeled the change in value from pre-testing to the post-acute period as:

$$(\mathrm{Y}_{\text{post-acute}} - \mathrm{Y}_{\text{pre-testing}} | Exposure, Covariates) = \beta_0 + \beta_{EXPOSURE} \times Exposure + \beta_{COVAR} \times Covariates \quad [3]$$

where $\mathrm{Y}_{\text{pre-testing}}$ is the median value from all outpatient measurements obtained within 180 days prior to SARS-CoV-2 testing and $\mathrm{Y}_{\text{post-acute}}$ is the median value from all outpatient measurements within 365 days after entering the post-acute phase. Comorbidities were ascertained using a phecode-based mapping of the Charlson comorbidities (Appendix C).[71] Secondary analyses were performed on demographic subgroups (stratified by sex and race), and timing of the new diagnoses (before or after 60 days following recovery. Sensitivity analyses were also performed to assess effects of our model assumptions for loss to follow up, length of EHR history, the threshold for "phenotype case", and bias from differences in baseline clinical variables. Differences in phenotype outcomes are reported as adjusted odds ratios (ORs), 95% confidence intervals (CIs) using Wald's method, and associated *p*-values. Differences in continuous outcomes are reported as group-wise adjusted mean difference and 95% confidence intervals. Statistical significance was set using a Bonferroni correction for number of independent tests. Additional details on model covariates and sensitivity analyses are provided in Appendix B. All analyses were performed using the R package *PheWAS*.[32]

Ethics, reporting statements, and role of funders

This study was conducted with approval from the Vanderbilt University Institutional Review Board (study approval numbers: #200512, #200731) under a waiver of informed consent. Patients were not directly contacted for the study. All patient data were abstracted from the EHR registry and maintained in accordance with institutional and federal privacy laws. The study was reported according to the Reporting of studies Conducted using Observational Routinely-collected health Data (RECORD) and Structured Template and Reporting Tool for Real World Evidence (STaRT-RWE).[72,73] The funding institutions and agencies had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; nor in the decision to submit the manuscript for publication.

Results

Study population

We identified 195,860 adults tested for SARS-CoV-2 at VUMC during the study period. We excluded 9,755 who had missing data on birth date or sex, reported a history of COVID-19 infection but never had a positive SARS-CoV-2 RT-PCR test at VUMC, or died before reaching the post-acute phase, leaving 186,105 adults in the primary cohort (Figure 2). Among these, 30,088 (16.2%) tested positive. Median age at initial test was 46 years (IQR 32–61), 57.1% were female, and 4,677 were pregnant around the time of SARS-CoV-2 testing. We followed patients in the EHR registry for a median 412 days (IQR 274–528) resulting in 199,407 person-years of observation after testing, with 113,198 (60.8%) having at least one follow-up visit in our system after recovery. Additional demographic and clinical characteristics of the study population are shown in Table 3 and Appendix D.

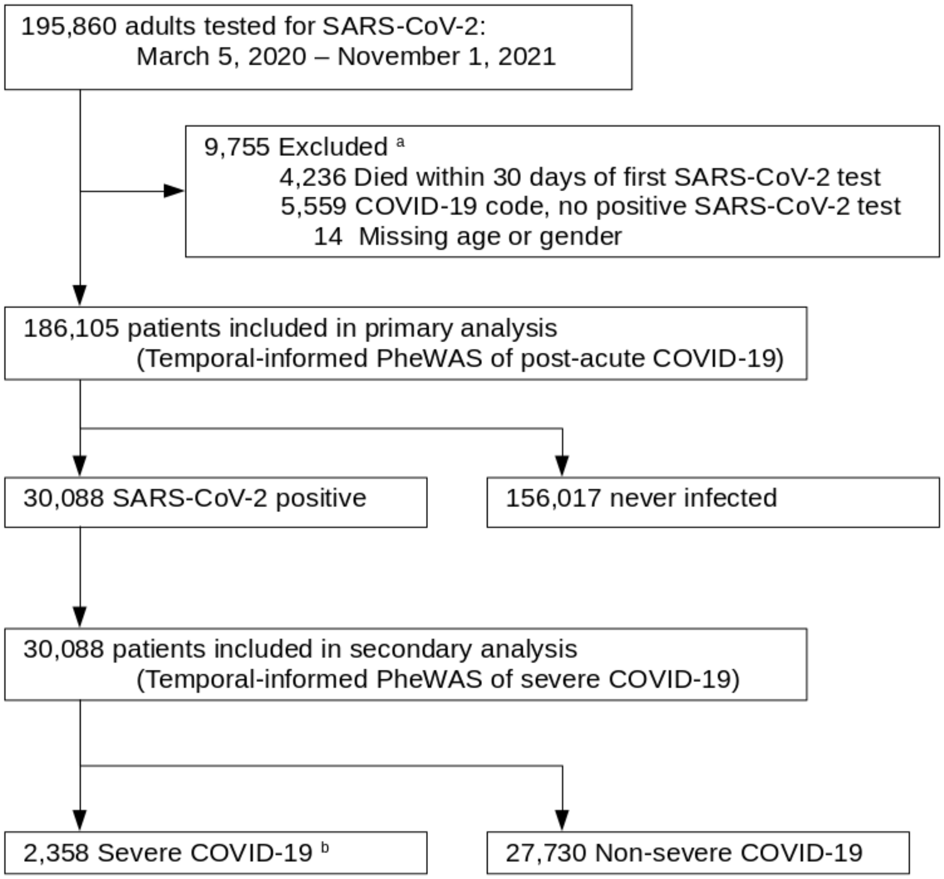

195,860 adults tested for SARS-CoV-2: March 5, 2020 – November 1, 2021
9,755 Excluded [a]
4,236 Died within 30 days of first SARS-CoV-2 test
5,559 COVID-19 code, no positive SARS-CoV-2 test
14 Missing age or gender
186,105 patients included in primary analysis (Temporal-informed PheWAS of post-acute COVID-19)
30,088 SARS-CoV-2 positive
156,017 never infected
30,088 patients included in secondary analysis (Temporal-informed PheWAS of severe COVID-19)
2,358 Severe COVID-19 [b]
27,730 Non-severe COVID-19

Figure 2. Study flow diagram

Flow diagram of adult patients in Vanderbilt COVID-19 EHR registry database, patients excluded, and numbers included in analyses. SARS-CoV-2: severe acute respiratory syndrome coronavirus 2. COVID-19: Coronavirus disease 2019.

[a] Some patients had more than one reason for exclusion.

[b] Severe COVID-19: admitted to hospital and required supplemental oxygen.

Table 3. Characteristics of registry cohort

Characteristic	Never Infected	SARS-CoV-2 Positive	Overall
Number in cohort	156,017	30,088	186,105
Age, median [IQR], years	46 [32, 62]	43 [30, 57]	46 [32, 62]
Sex (%)			
Female	89,547 (57.4)	16,718 (55.6)	106,265 (57.1)
Male	66,470 (42.6)	13,370 (44.4)	79,840 (42.9)
Race (%)			
Black	17,106 (11.0)	3,274 (10.9)	20,380 (11.0)
Other race or multiracial	7,901 (5.1)	1,714 (5.7)	9,615 (5.2)
Unknown/not reported	18,996 (12.2)	5,924 (19.7)	24,920 (13.4)
White	112,014 (71.8)	19,176 (63.7)	131,190 (70.5)
Ethnicity (%)			
Hispanic/Latino	4,759 (3.1)	1,217 (4.0)	5,976 (3.2)
Non-Hispanic/Non-Latino	128,049 (82.1)	21,936 (72.9)	149,985 (80.6)
Unknown/not reported	23,209 (14.9)	6,935 (23.0)	30,144 (16.2)
Received care at VUMC prior to SARS-CoV-2 test (%) [a]	106,839 (68.5)	20,860 (69.3)	127,699 (68.6)
SARS-CoV-2 testing indication (%)			
Asymptomatic screening [b]	89,727 (57.5)	6,095 (20.3)	95,822 (51.5)
Symptomatic testing	66,290 (42.5)	23,993 (79.7)	90,283 (48.5)
EHR observation time			
After SARS-CoV-2 test, median [IQR], days	420 [267, 533]	392 [317, 459]	412 [274, 528]
After recovery, median [IQR], days	378 [215, 495]	361 [285, 427]	374 [224, 489]
Hospitalization associated with SARS-CoV-2 test (%) [c]	43,146 (27.7)	3,393 (11.3)	46,539 (25.0)
Severe COVID-19 (%) [d]	-	2,358 (7.8)	-
Follow-up visit type (%) [e]			
Any follow-up visit	96,615 (61.9)	16,583 (55.1)	113,198 (60.8)
Office visit	89,559 (57.4)	15,593 (51.8)	105,152 (56.5)
Laboratory / anti-coagulation visit	42,646 (27.3)	7,216 (24.0)	49,862 (26.8)
Inpatient surgery or procedure	27,213 (17.4)	4,091 (13.6)	31,304 (16.8)
Telemedicine visit	16,617 (10.7)	2,478 (8.2)	19,095 (10.3)
Outpatient surgery or procedure	19,725 (12.6)	2,728 (9.1)	22,453 (12.1)
Allied health practitioner visit [f]	14,821 (9.5)	2,580 (8.6)	17,401 (9.4)
Infusion / radiation care	4,043 (2.6)	542 (1.8)	4,585 (2.5)
Maternity care	3,899 (2.5)	482 (1.6)	4,381 (2.4)
Outpatient observation in Emergency Department	2,403 (1.5)	422 (1.4)	2,825 (1.5)
Inpatient medical admission	1,197 (0.8)	1,239 (4.1)	2,436 (1.3)
Time from SARS-CoV-2 test to first follow-up visit, median [IQR], days	66 [44, 139]	86 [48, 181]	69 [44, 145]
Pregnant during study observation period (%)	7,565 (4.8)	609 (2.0)	8,174 (4.4)
Pregnant around time of SARS-CoV-2 test (%)	4,488 (2.9)	189 (0.6)	4,677 (2.5)
Died during post-acute phase (%)	1,535 (1.0)	158 (0.5)	1,693 (0.9)

[a] Defined as having at least two visits at VUMC prior to SARS-CoV-2 test separated by ≥180 days.

[b] Reasons for asymptomatic screening included: asymptomatic admission to the hospital for another diagnosis, pre-procedural or pre-surgical screening, known SARS-CoV-2 exposure, pre-receipt of immunosupressive or anti-neoplastic therapy, pre-transplant evaluation, or requirement for placement in post-acute care or long-term nursing care.

[c] SARS-CoV-2 test performed within 15 days prior to a hospital admission or during a hospital admission.

[d] Severe COVID-19: admitted to hospital and received supplemental oxygen.

[e] Some patients had more than one visit type.

[f] Allied health practitioner visits included visits coded as being nurse-only visits, dietitian or nutritionist visits, and clinical support or educational visits.

Temporal-informed phenotyping of post-acute period

At the data censoring date and after mapping for diagnosis-specific exclusions, 1,347 phecodes were well-represented in the study population with ≥10 phenotype cases under the naive approach. Most diagnosis codes entered in the EHR after recovery pertained to conditions that were also present before the post-acute phase. After applying our temporal-informed phenotyping to identify new diagnoses following recovery, the median case retention per phecode was 36.1% (IQR: 23.6% – 51.5%) and 902 (70.0%) phecodes remained well-represented in the cohort. Figure 3 illustrates the distribution of case retention by phecode chapter. Phenotypes in the musculoskeletal, dermatologic, and symptoms chapters were most likely to represent new diagnoses in the post-acute period, whereas neoplasms were least likely to represent new diagnoses (Appendix E).

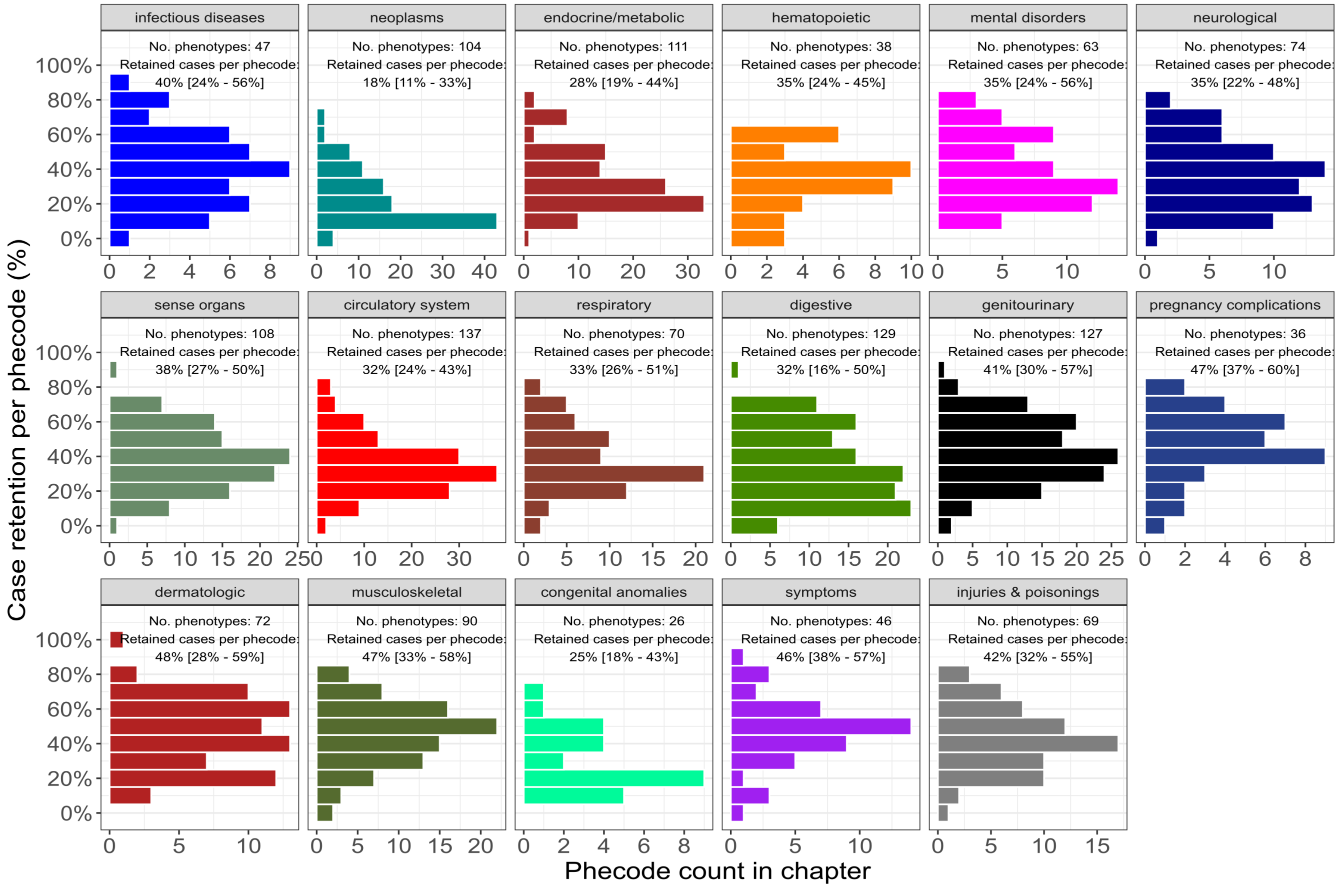


Figure 3. Phecode case retention by temporal-informed phenotyping.

Histograms of phenotype case retention per PheWAS code (phecode) using temporal-informed phenotyping. Individual histograms indicate each chapter within the phecode hierarchy.[24] Number of phecodes per chapter are shown on *x* axis, case retention per phecode is shown on *y* axis,. Labels indicate number of phenotypes with ≥10 cases and median [interquartile range] of the per-phecode case retention in each chapter.

Control retention under temporal-informed phenotyping was high (per-phecode median 91.7%; IQR: 87.9% - 95.1%; Figure 4), although several respiratory phenotypes (e.g. shortness of breath, cough, abnormal chest sounds) had lower control retention as these phecodes were very common around the date of testing for SARS-CoV-2. Patients with ≥6 months of care at VUMC prior to testing were more likely to have at least one new diagnoses in the EHR under temporal-informed phenotyping compared to patients with no substantial care history at our institution (39.1% vs. 30.8%, $p < 1.0\times10^{-15}$), indicating the temporal-informed phenotypes were not driven by patients with short EHR histories.

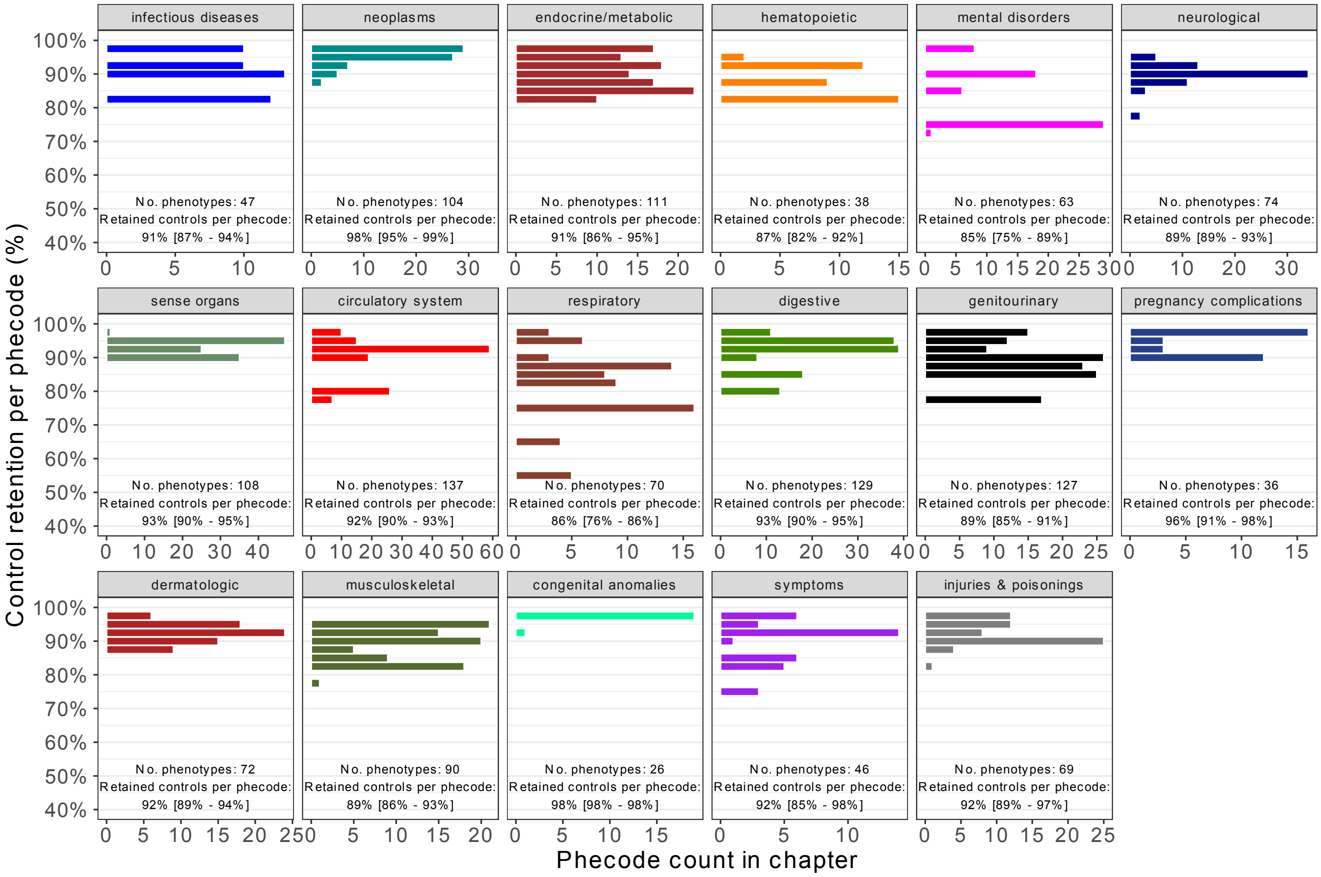


Figure 4. Control retention by temporal-informed phenotyping

Histograms of phenotype control retention per PheWAS code (phecode) using temporal-informed phenotyping. Individual histograms indicate each chapter within the phecode hierarchy.[22] Number of phecodes per chapter are shown on *x* axis, control retention per phecode is shown on *y* axis,. Labels indicate number of phenotypes with ≥10 cases and median [interquartile range] of the per-phecode case retention in each chapter.

Temporal-informed PheWAS identifies new post-acute phenotypes in COVID-19 survivors

Temporal-informed PheWAS demonstrated that survivors of COVID-19 had increased odds for developing 43 distinct phenotypes during outpatient follow-up (Figure 5, Table 4). Phenotypes that reached phenome-wide significance encompassed 12 disease categories, with circulatory (7 phenotypes), pregnancy complications (7 phenotypes), respiratory (5 phenotypes), and neurological (4 phenotypes) chapters having the greatest number of associated phenotypes.

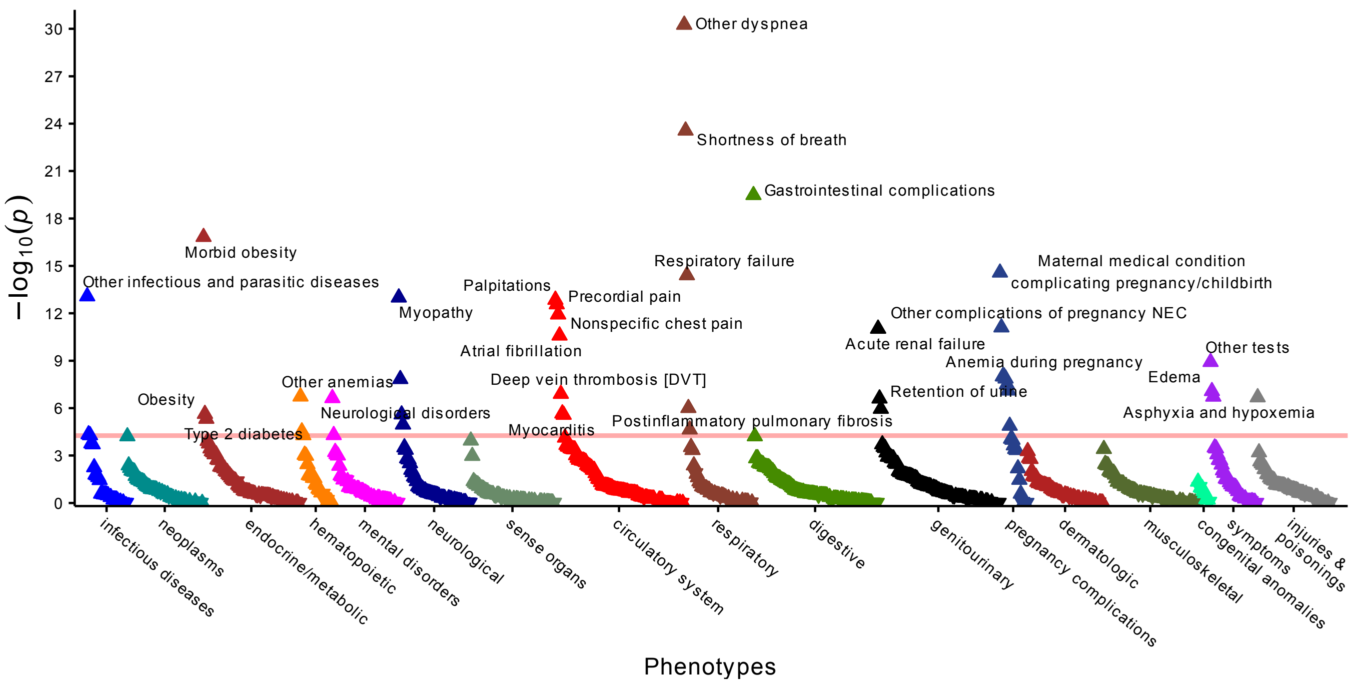


Figure 5. Temporal-informed phenome scan of post-acute COVID-19.

PheWAS plot of new post-acute phenotypes identified by temporal-informed phenotyping for COVID-19 survivors versus never-infected patients as the referent group (n = 186,105, phenotypes available for testing = 902). The *x* axis represents phecodes grouped by chapter within the phecode hierarchy. The *y* axis represents the negative log-transformed *p* values obtained using logistic regression after adjusting for age, sex, race, ethnicity, length of EHR observation after recovery, indication for testing, and medical comorbidities prior to testing. Upward triangles represent phenotypes with odds ratio >1.0 for COVID-19 survivors and downward triangles represent phenotypes with odds ratio <1.0. Horizontal red line indicates the phenome-wide significance *p* value significance using a Bonferroni correction (p=5.54×10^{-5}).

Table 4. Summary of temporal-informed PheWAS in post-acute COVID-19

Phecode[a]	Description	Odds Ratio	95% CI	p value	No. cases	No. controls
512.9	Other dyspnea	3.04	(2.52-3.68)	5.54×10^{-31}	811	93,936
512.7	Shortness of breath	2.49	(2.09-2.96)	2.73×10^{-24}	988	93,936
569.2	Gastrointestinal complications of surgery	6.54	(4.38-9.75)	3.32×10^{-20}	116	166,825
278.11	Morbid obesity	2.35	(1.93-2.86)	1.49×10^{-17}	624	154,861
649	Conditions of the mother complicating pregnancy, childbirth, or the puerperium	3.85	(2.76-5.38)	2.66×10^{-15}	169	95,518
509.1	Respiratory failure	7.09	(4.35-11.6)	3.89×10^{-15}	101	157,792
136	Other infectious and parasitic diseases	9.20	(5.14-16.5)	8.43×10^{-14}	54	181,966
359.2	Myopathy	20.5	(9.24-45.4)	9.99×10^{-14}	33	174,863
427.9	Palpitations	2.14	(1.75-2.61)	1.40×10^{-13}	628	137,086
418.1	Precordial pain	3.21	(2.35-4.39)	2.71×10^{-13}	278	138,537
418	Nonspecific chest pain	2.01	(1.66-2.43)	1.19×10^{-12}	746	138,537
646	Other complications of pregnancy NEC	5.91	(3.55-9.83)	7.89×10^{-12}	69	99,542
585.1	Acute renal failure	3.15	(2.26-4.38)	9.49×10^{-12}	309	157,475
427.21	Atrial fibrillation	2.62	(1.98-3.48)	2.56×10^{-11}	443	137,086
1010	Other tests	3.17	(2.19-4.60)	1.21×10^{-9}	155	169,347
644	Anemia during pregnancy	7.43	(3.74-14.7)	9.91×10^{-9}	38	101,761
1010.6	Reproductive and maternal health services	1.75	(1.44-2.12)	9.99×10^{-9}	591	172,787
638	Other high-risk pregnancy	2.19	(1.67-2.86)	1.34×10^{-8}	312	178,757
350.1	Abnormal involuntary movements	2.53	(1.83-3.48)	1.46×10^{-8}	256	170,487
671	Venous/cerebrovascular complications & embolism in pregnancy and the puerperium	21.5	(7.25-63.7)	3.10×10^{-8}	17	103,586
649.1	Diabetes or abnormal glucose tolerance complicating pregnancy	4.73	(2.68-8.34)	7.77×10^{-8}	57	95,518
782.3	Edema	2.08	(1.59-2.73)	8.34×10^{-8}	424	168,184
452.2	Deep vein thrombosis [DVT]	3.23	(2.09-4.99)	1.26×10^{-7}	138	162,711
285	Other anemias	2.05	(1.56-2.68)	1.85×10^{-7}	473	146,505
781	Symptoms involving nervous and musculoskeletal systems	3.07	(2.01-4.68)	1.88×10^{-7}	151	180,070
1013	Asphyxia and hypoxemia	5.51	(2.89-10.5)	2.07×10^{-7}	52	175,439
292	Neurological deficits	2.39	(1.72-3.32)	2.31×10^{-7}	242	162,234
599.2	Retention of urine	2.93	(1.95-4.41)	2.45×10^{-7}	184	149,134
514	Abnormal findings examination of lungs	2.29	(1.64-3.20)	9.86×10^{-7}	350	163,569
587	Kidney replaced by transplant	32.4	(7.99-131.)	1.12×10^{-6}	22	157,475
401.1	Essential hypertension	1.42	(1.23-1.64)	2.17×10^{-6}	1,698	122,907
278.1	Obesity	1.70	(1.36-2.12)	2.33×10^{-6}	566	154,861
327.32	Obstructive sleep apnea	1.69	(1.36-2.11)	2.51×10^{-6}	669	150,608
420.1	Myocarditis	10.0	(3.83-26.2)	2.67×10^{-6}	20	177,003
250.2	Type 2 diabetes	1.77	(1.38-2.25)	4.75×10^{-6}	572	148,033
348.8	Encephalopathy, not elsewhere classified	6.23	(2.76-14.1)	1.10×10^{-5}	32	160,519
653	Problems associated with amniotic cavity and membranes	8.04	(3.15-20.5)	1.32×10^{-5}	19	97,532
502	Post-inflammatory pulmonary fibrosis	5.47	(2.49-12.0)	2.26×10^{-5}	40	157,792
284.1	Pancytopenia	3.25	(1.87-5.66)	2.96×10^{-5}	94	146,505
38.3	Bacteremia	8.03	(2.95-21.9)	4.54×10^{-5}	19	166,009
292.3	Memory loss	1.99	(1.43-2.77)	5.09×10^{-5}	287	162,234
285.21	Anemia in chronic kidney disease	3.10	(1.79-5.36)	5.22×10^{-5}	104	146,505
54	Herpes simplex	3.66	(1.95-6.85)	5.22×10^{-5}	54	149,827

[a] A list of ICD-10-CM codes included in each phecode is available at: https://phewascatalog.org/phecodes_icd10cm [22]

In contrast, the naive approach identified 219 phenotypes reaching Bonferroni-adjusted significance (Appendix F). Although the top associations by temporal-informed phenotyping were also observed in the naive analysis, discerning the clinical relevance of any association in the naive analyses was difficult due to the high number of associations pertaining to phenotypes of acute illness (e.g. altered mental status, hypotension, respiratory failure, sepsis, septicemia, acidosis) or chronic medical conditions know to be risk factors for COVID-19 (e.g. chronic kidney disease, essential hypertension, hyperlipidemia).[64,65] Only 28 phenotypes identified by temporal-informed phenotyping were found among the top 100 diagnoses identified by naive phenotyping. Additionally, associations with phenotypes for memory loss and post-inflammatory pulmonary fibrosis were only seen using temporal-informed analyses. Strength of associations (based on p-value) were higher under the naive approach due to higher phenotype case counts, but adjusted odds ratios were similar under both approaches (Figure 6).

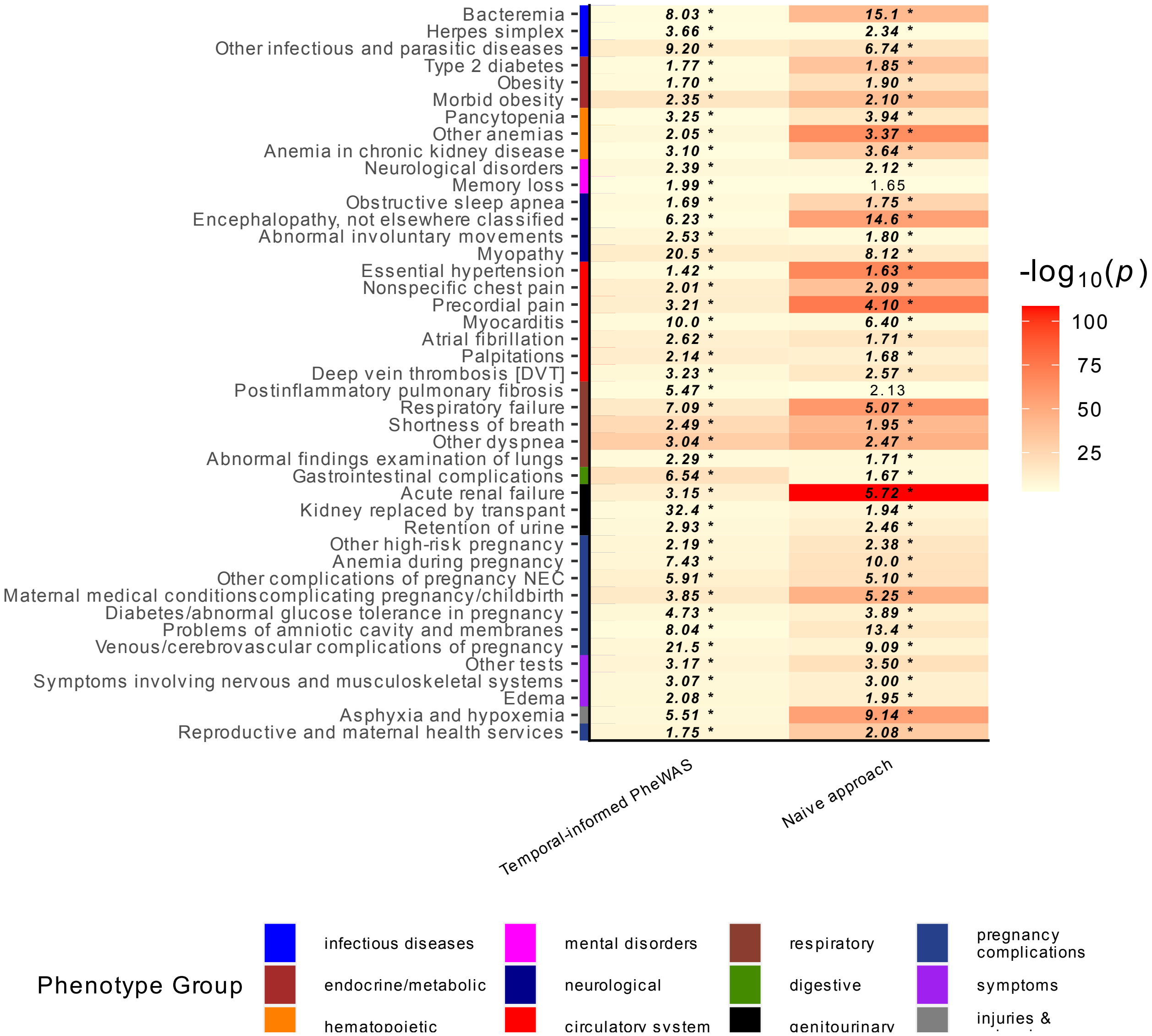


Figure 6. Comparison of PheWAS results by temporal-informed or naive phenotyping.

Comparison of PheWAS results using temporal-informed phenotyping (left column) and naive post-acute phenotyping (right column). The *y* axis represents phenotypes (as PheWAS codes / "phecodes") that are group by category within the phecode hierarchy.[24] Cell color intensity illustrates adjusted *p* values by logistic regression. Text in cells show point estimates for effect odds ratios. Text in bold/italic and with a '*' indicate PheWAS associations that were statistically significant using a Bonferroni correction. Results for phecodes that were significant in the primary analysis (left column) are displayed for brevity. Results in bold/italic text and with a '*' indicate PheWAS associations that were statistically significant using a Bonferroni-corrected *p* value.

Figure 7 illustrates subgroup analyses based on demographics and timing of the post-acute diagnoses. New post-acute phenotypes related to gastrointestinal complications of surgery, obesity, abnormal glucose control, pregnancy complications, and anemia were common to both White, Non-Hispanic and Black, Non-Hispanic subgroups, while new chronic fatigue syndrome was unique among Black, Non-Hispanic COVID-19 survivors. Phenotypic associations were evenly distributed among males and females, although males had more phenotypes related to new abnormal pulmonary function while females had more new cardiovascular phenotypes. Many of the temporal-informed diagnoses were initially made late (>60 days) into the post-acute period, however 14 phenotypes presented earlier during the first 60 days after recovery. Subgroup PheWAS results are available in the Appendix G. Our findings were also robust to several sensitivity analyses. Most phenotypic associations were replicated when using 1) patients with ≥1 follow-up visit in our system after recovery, 2) patients with an EHR length ≥6 months prior to testing, 3) using a less stringent phenotype case threshold, and 4) a propensity-matched cohort which matched 3 never-infected controls to each COVID-19 survivor (Appendix H).

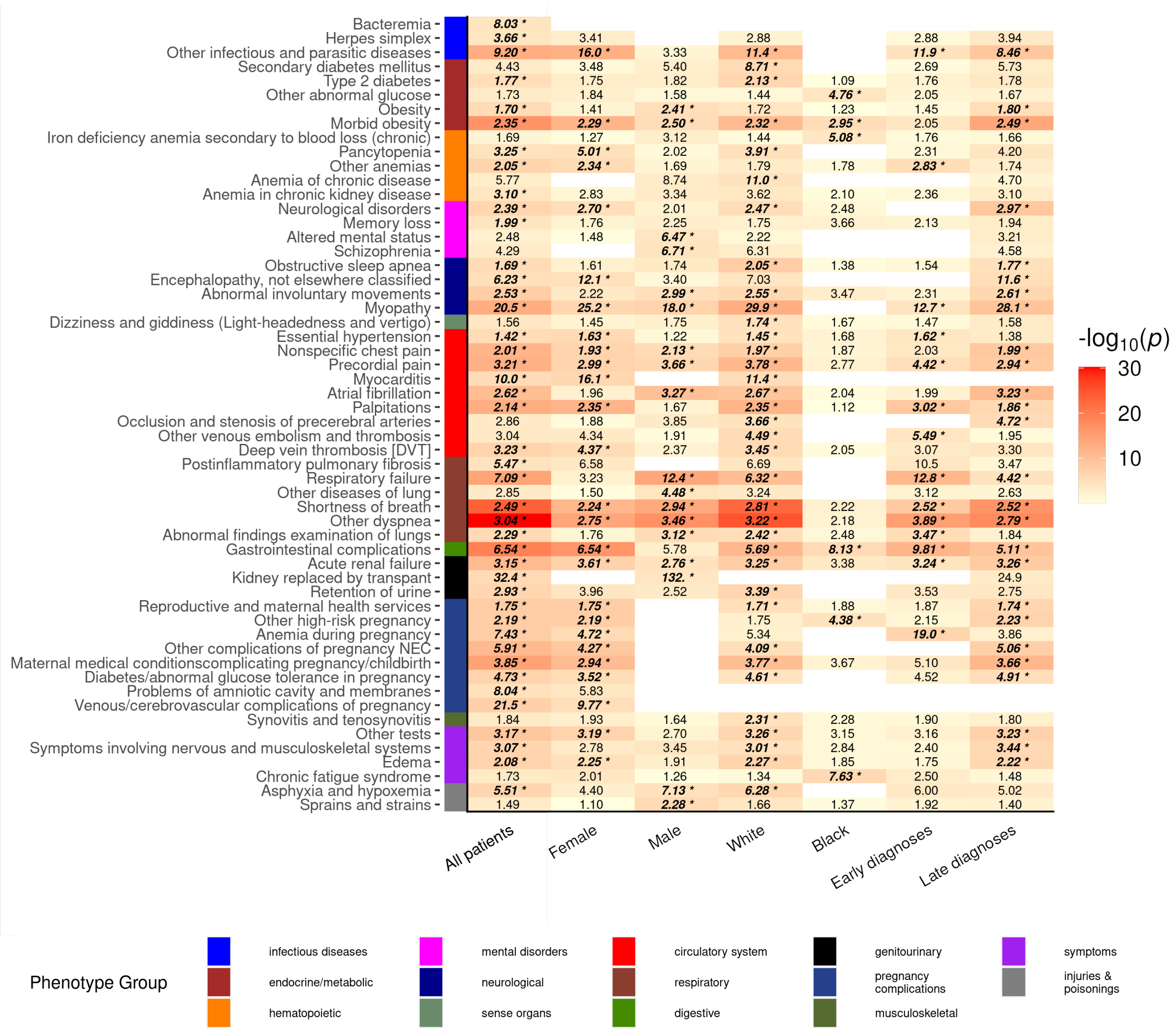


Figure 7. Temporal-informed phenome scans of post-acute COVID-19 by demographic subgroups and timing of post-acute diagnoses.

PheWAS results for new post-acute phenotypes identified by temporal-informed phenotyping among all adults tested for SARS-CoV-2 (left column, n=186,105), stratified by demographic subgroups (male sex, female sex, white non-Hispanic, black non-Hispanic), and stratified by onset of the new diagnoses ("Early" diagnoses: within 60 days after recovery; "Late diagnoses": later than 60 days after recovery). The *y* axis represents phecodes group by chapter within the phecode hierarchy. Cell color intensity illustrates adjusted *p* values by logistic regression. Text in cells show point estimates for effect odds ratios. Text in bold/italic and with a '*' indicate PheWAS associations that were statistically significant using a Bonferroni correction. Results for phecodes with a statistically significant association in any subgroup analysis are displayed. Empty cells indicate analyses with insufficient phenotype cases (less than 10) to perform the analysis for that phenotype in the subgroup.

Post-acute clinical phenotypes associated with severe COVID-19

Among the 30,088 COVID-19 survivors, those with severe disease (n=2,358, 7.8%) had substantially higher odds of developing multiple respiratory and cardiovascular phenotypes with the top phenotypic associations being new respiratory failure, hypertension, and abnormalities on lung examination. Additional post-acute phenotypes associated with severe SARS-CoV-2 survivors are shown in Table 5.

Table 5. Summary temporal-informed PheWAS for severe COVID-19 survivors

Phecode[a]	Description	Odds Ratio	95% CI	p value	No. cases	No. controls
509.1	Respiratory failure	225	(62.7-808)	1.02×10^{-15}	31	25,204
401.1	Essential hypertension	3.71	(2.55-5.39)	6.72×10^{-12}	243	21,801
514	Abnormal findings examination of lungs	10.7	(4.93-23.4)	2.30×10^{-9}	42	25,588
504	Other interstitial lung disease	142	(24.7-818)	1.55×10^{-6}	10	25,204
507	Pleurisy or pleural effusion	28.5	(7.92-103)	1.76×10^{-6}	14	25,204
427.21	Atrial fibrillation	4.26	(2.38-7.63)	6.11×10^{-6}	68	23,263
798	Malaise and fatigue	2.91	(1.87-4.52)	1.95×10^{-6}	162	19,803
276.13	Hyperpotassemia	12.0	(4.15-34.7)	4.45×10^{-6}	24	24,600
502	Post-inflammatory pulmonary fibrosis	47.5	(8.11-278)	1.86×10^{-5}	10	25,204
250.22	Type 2 diabetes with renal manifestations	45.7	(7.79-268)	2.30×10^{-5}	32	24,221
1013	Asphyxia and hypoxia	11.8	(3.45-40.5)	8.59×10^{-5}	15	26,963

[a] A list of ICD-10-CM codes included in each phecode is available at: https://phewascatalog.org/phecodes_icd10cm [28]

Validation of select temporal-informed phenotypic associations in the EHR

As several phenotypes identified in our temporal-informed analyses are ostensibly chronic conditions, we selected a subset of the temporal-informed phenotypic associations that had structured EHR data readily available via an associated vital sign or laboratory test (e.g. body mass index [BMI] for obesity, blood pressure for hypertension, hemoglobin level for anemia). We then assessed if SARS-CoV-2 infection was also associated with changes in the vital sign or lab value from pre-testing to post-acute periods among patients with normal values prior to SARS-CoV-2 testing. As an example, among the 37,838 patients who were not obese (BMI < 30) and had both pre-testing and post-acute BMI recorded in the EHR, BMI increased by 0.21 (±1.4) kg/m^2 in COVID-19 survivors compared to 0.01

(±1.6) kg/m^2 in never infected patients (adjusted mean difference: 0.16; 95% CI: 0.12-0.21; p=2.00×10^{-13}). COVID-19 survivors also tended to have more substantial changes in heart rate and white blood cell (WBC) count, compared to never infected patients (Table 6, Figure 8). Small changes were also noted in systolic blood pressure, respiratory rate, and estimated glomerular filtration rate although difference for these values were smaller than the minimum unit of measure for these variables. Although these differences between groups were small (~1-2% of typical baseline values) the vital sign changes aligned with the direction of the associated clinical phenotype. We did not observe substantial differences between groups in labs for hemoglobin, platelets, serum potassium, hemoglobin A1C, or serum glucose.

Table 6. Changes in outpatient vital signs or laboratory studies for select temporal-informed phenotypes

			Change in lab or vital sign from pre-testing to post-acute[a]			
Post-acute phenotype(s)	**Vital sign / Lab (units)**	**Subgroup[b]**	**Never Infected mean (SD)[c]**	**SARS-CoV-2 Positive mean (SD)[c]**	**Adjusted mean difference (95% CI)[d]**	***p* value[e]**
Obesity Morbid obesity	BMI (kg/m^2)	Non-obese (n=37,838)	0.01 (1.6)	0.21 (1.4)	0.16 (0.12 - 0.21)	2.00 × 10^{-13}
Essential hypertension	Systolic Blood Pressure (mmHg)	Normal blood pressure or pre-hypertension (n=28,912)	-0.2 (13.0)	0.4 (12.0)	0.5 (0.1 - 1.0)	0.015
Palpitations Atrial fibrillation	Heart Rate (bpm)	Normal heart rate, no arrhythmia diagnoses (n=31,364)	0.1 (12)	1.1 (12)	1.0 (0.6 - 1.3)	3.81 × 10^{-7}
Respiratory failure	Respiratory Rate (min^{-1})	Normal respiratory rate, no lung disorders (n=19,764)	-0.1 (2.2)	0.1 (2.3)	0.2 (0.1 - 0.3)	3.89 × 10^{-5}
Pancytopenia	White Blood Cell (10^3/µL)	Normal WBC, no hematologic disorders (n=12,346)	0.0 (1.9)	0.2 (1.9)	0.2 (0.1 - 0.3)	5.72 × 10^{-6}
Acute renal failure	Estimated GFR (ml/min)	No renal failure or kidney transplant (n=14,305)	0 (13)	1 (12)	1 (0 - 1)	0.008

[a] Among patients with the vital sign or lab value recorded both within 180 days prior to SARS-CoV2 testing and within 365 days following recovery.

[b] Prior to SARS-CoV-2 testing

[c] Calculated for each patient as Y$_{\text{post-acute}}$ - Y$_{\text{pre-testing}}$, where Y is the vital sign value or laboratory value. Negative values indicate a decrease in the vital sign / lab value from the pre-testing to the post-acute phases, and positive values indicate an increase in the vital sign / lab value.

[d] Mean difference and 95% confidence interval between groups adjusted for age, sex, race, ethnicity, and time between pre-SARS-CoV-2 test value and post-acute value.

[e] Adjusted p-values using linear regression

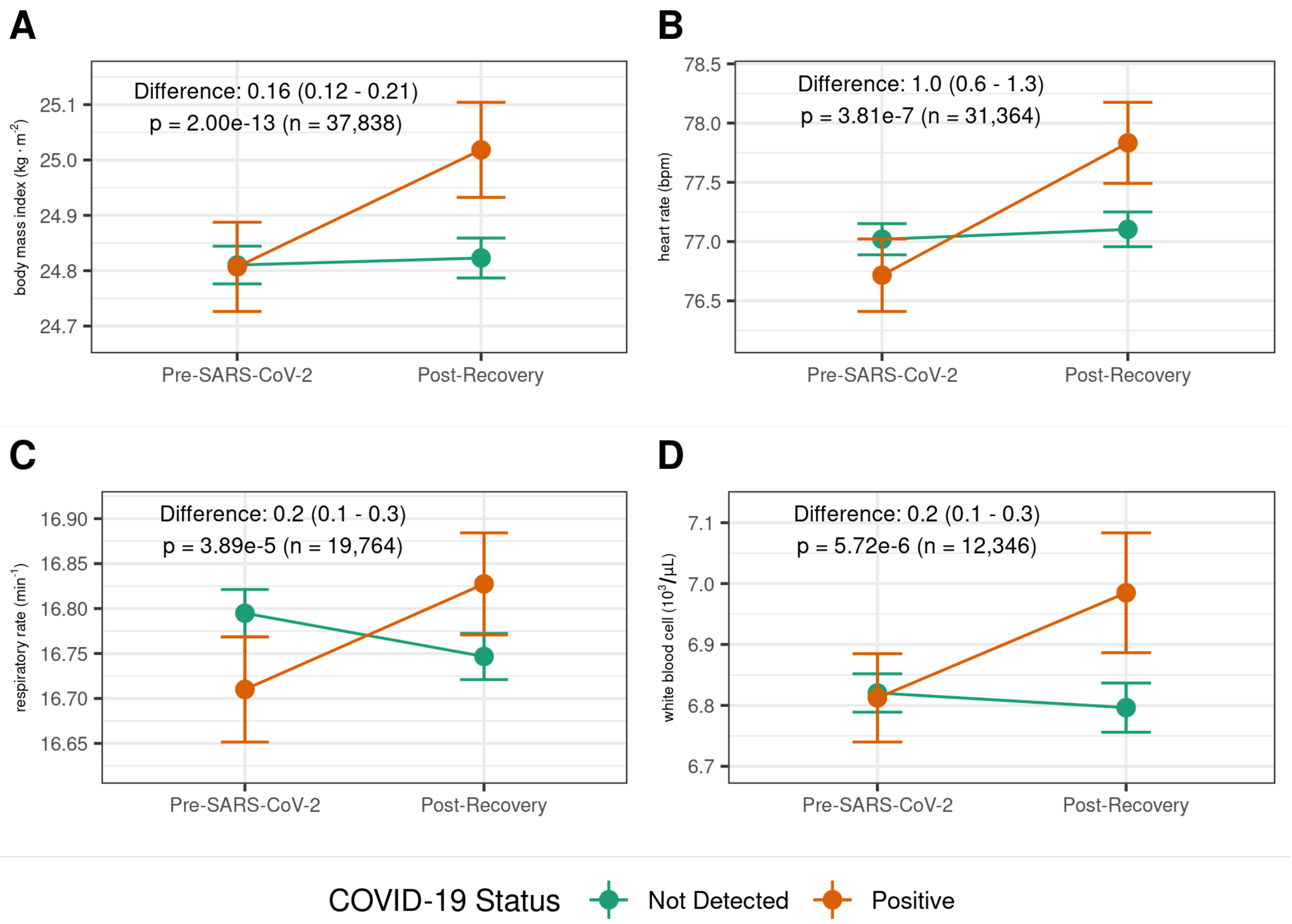


Figure 8. Changes in select vital signs and laboratory test values in post-acute COVID-19.

COVID-19 survivors (orange) had more substantial changes in **A.** body mass index, **B.** heart rate, **C.** respiratory rate, and **D.** white blood cell count from pre-testing to post-recovery compared with never-infected controls (green). For each patient we used the median pre-testing values obtained during outpatient visits occurring within 180 days before the index SARS-CoV-2 test, and the median post-recovery values obtained during outpatient visits occurring within 365 days after recovery from illness. Dots represent mean values in each exposure group, bars represent standard errors of the mean. Labels represent the adjusted mean difference between COVID-19 survivors and never-infected controls, number of patients with data for each analysis, and *p*-values obtained by multiple linear regression.

Discussion

Principal findings

Temporal-informed phenotyping identified a range of new diagnoses among COVID-19 survivors affecting multiple organ systems. Compared with the naive approach of using all diagnosis codes occurring after the event, temporal-informed phenotyping was less influenced by phenotypes related to acute illness or previous medical history. While the underlying mechanisms of these post-acute manifestations of COVID-19 remain uncertain, they may reflect late effects of inflammation or vascular injury and the sequelae of severe illness among hospitalized survivors.[38,39] Several post-acute phenotype associations were also supported by changes in vital signs values from pre-testing to the post-acute period. Although the observed differences in vital signs attributable to COVID-19 survivorship were typically small, they still may have substantial long-term implications on a population-level scale. A meta-analysis of 46 prospective cohort studies found an increase in resting heart rate by 10 bpm was associated with a 9% increase in all-cause mortality and 8% increase in cardiovascular mortality.[74] Thus, given the unprecedented scale of the COVID-19 pandemic, even the modest changes in these parameters observed in our study may portend profound long-term implications on public health.

Comparison with other studies

Our findings align with other reports on long-term consequences of COVID-19.[38,40–50] Ayoubkhani et.al. found increased rates of death, hospital readmission, diabetes, cardiovascular events, and chronic kidney and liver disease among COVID-19 survivors using hospital administrative data from the United Kingdom.[41] Daugherty et. al. observed increased risk of multiple new cardiovascular, respiratory, hematologic, and neurologic diagnoses among COVID-19 survivors using insurance administrative claims data from the United States.[45] Al-Aly et. al. reported excess burden of respiratory, nervous system, metabolic, mental health, cardiovascular, and gastrointestinal disorders among COVID-19 survivors receiving care through the US Veterans Health Administration.[47] Similar to our findings of increased myopathy, neurological deficits, encephalopathy, and memory loss, Tacquet et. al. found that COVID-19 survivors had elevated risk for developing multiple neurologic and psychiatric disorders in a multinational EHR dataset.[48] Estiri et al. evaluated the temporal evolution post-acute COVID-19 phenotypes among patients in a single US academic center using a sequence-based framework MLHO, also observing substantially increased rates of cardiovascular, respiratory, endocrine, and neurologic phenotypes among COVID-19 survivors.[51]

Strengths

Our temporal-informed phenotyping framework naturally augments classical PheWAS, allowing us to identify potential post-acute sequelae of COVID-19 and replicate several associations identified in other studies. The distribution of case retention under temporal-informed phenotyping for various phecodes aligned with our clinical experience. Phecode chapters with more short-lived conditions like symptoms, musculoskeletal, and dermatologic diagnoses had the highest case retention, while chapters with mostly chronic diagnoses such as neoplasms and congenital abnormalities had the lowest case retention. Although other phenotyping approaches incorporating temporal information have been reported, many rely upon complex machine learning methods that require specialized computational expertise, and/or focus on predicting a specific disease processes or future outcome.[17,25,51,75–77] In contrast, our method uses PheWAS in a hypothesis-free approach to broadly scan the entire medical phenome for new diagnoses occurring at any time after a discrete medical event. The PheWAS framework has several advantages over other high-throughput phenotyping approaches. It reduces the phenome feature space size from ~80,000 ICD-10-CM codes to ~1800 clinically relevant phecodes, improving computational efficiency. The phenotype feature engineering method in the *PheWAS* software package automatically incorporates diagnosis-specific exclusion criteria to limit contamination of controls with potential cases, providing additional specificity compared to other phenotyping methodologies.[11,24,32] PheWAS analyses are also more accessible to researchers than more complex machine learning methods.[78] Thus our temporal-informed phenotyping could be easily adapted to examine the post-acute phenotype consequences among survivors other acute medical event such as pneumonia or sepsis.[79,80]

VUMC is a major provider of primary through quaternary care in the American Mid-South and encompasses a broad patient population seeking SARS-CoV-2 testing. Follow-up rates were relatively high with 113,198 (60.8%) patients having at least one follow-up visit in the post-acute phase. This study leveraged our longstanding institutional experience with using the EHR for secondary research,[4,11] allowing us to capture deep phenotyping information, such as SARS-CoV-2 testing indication and setting of post-acute diagnoses, which may not be well-represented in administrative datasets or cross-institutional research databases.[47,81] We were also able to compare temporal-informed phenotypes between survivors of severe COVID-19 versus survivors of non-severe COVID-19, and we correlated several temporal-informed phenotypic associations with changes in vital signs or laboratory values from the pre-testing to post-acute periods.

Limitations

As with all observational studies, residual confounding is possible as not all relevant risk factors for COVID-19 are well-represented in the EHR (e.g. social interactions, household members, or travel history), but we included a broad set of clinical and EHR covariates in our PheWAS models that are available in many EHRs. We used in-house SARS-CoV-2 test results to identify COVID-19 cases which may have a higher sensitivity than diagnostic billing codes,[67,82] but not all regional clinics / hospitals share our EHR and some of our “never infected” patients may have tested positive elsewhere. To mitigate risk of misclassifying COVID-19 status we excluded all patients who reported a clinical diagnosis of COVID-19 but did not have a corresponding positive PCR test in our EHR. Additionally, patients in our study may have received post-acute care at outside facilities; those diagnoses that may not have been available in our EHR. Given the highly fragmented nature of the US healthcare system, this data fragmentation risk is inherent to any US study using real-world EHR data. Our institution mostly draws patients from the American Mid-South, thus our findings may not be generalizable to other patient populations, but we anticipate extending this methodology to larger multicenter networks in future work. Although ICD-coded diagnoses are commonly used in EHR cohort studies, they may not fully describe the spectrum of symptoms reported by COVID-19 survivors, and additional analyses examining symptoms and clinical findings extracted from narrative text could reveal additional disease patterns in this population.[17] This study also did not examine differences among survivors of various SARS-CoV-2 variants as variant typing is not routinely performed at our institution. The B.1.1.7-Alpha variant was the dominant strain in Tennessee until early July 2021, with the B.1.617.2-Delta variant remaining dominant through the remainder of the observation period.[83] Additional analyses will be necessary in the future to assess how novel SARS-CoV-2 variants including BA.1-Omicron may influence long-term outcomes among COVID-19 survivors in our region. Finally, our study design can only detect clinical associations between COVID-19 and development of new medical phenotypes; further studies are required to understand the mechanisms underlying these disease associations.

Conclusion

Temporal-informed phenotyping naturally augments the traditional PheWAS framework. Using temporal-informed PheWAS, we found that COVID-19 survivors in our institutional EHR registry had increased risk for a broad range of new medical problems after recovery from acute illness. PheWAS with temporal-informed phenotyping represents a promising approach to study the phenotypic consequences of acute medical conditions like COVID-19 over time, enabling rapid assessment of the entire medical phenome at population-level scales. These findings can assist clinicians in identifying medical problems arising among survivors of acute medical events, allow researchers to efficiently coordinate studies of morbidity trends, and help policymakers plan for the ongoing health consequences of future pandemics.

CHAPTER 3

SUMMARY

This study describes the development of a temporal-informed phenotyping framework that incorporates information on prior diagnoses to identify temporal changes in patients' medical phenome. We show that temporal-informed phenotyping can differentiate phecode chapters with more transient or temporary diagnoses (e.g. musculoskeletal, dermatologic, and symptoms phenotypes) from to phecode chapters consisting of mostly chronic conditions (e.g. neoplasms and congenital abnormalities), and show that many phenotypes across the medical phenome had substantial rates of being new after a COVID-19 test. We then demonstrate the utility of temporal-informed phenotyping to identify new diagnoses occurring among COVID-19 survivors using a large EHR cohort of patients tested for the disease at Vanderbilt University Medical Center, identifying 43 new phenotypes associated with COVID-19 survivorship. Compared to a naive post-event phenotyping approach, temporal-informed phenotyping provides more interpretable findings by better excluding phenotypes associated with acute illness or prior comorbidities. We further demonstrate the robustness of our findings in subgroup and sensitivity analyses, and provide support for several observed post-COVID phenotypic associations by identifying concomitant changes in associated vital signs and laboratory test results from pre-testing to the post-recovery periods among patients with normal values prior to contracting COVID-19.

Limitations

Several limitations of our temporal-informed phenotyping merit discussion as it is currently designed. Firstly, we performed the study using EHR data from a single academic medical center, which may impact the generalizability of our results. Secondly, all limitations of traditional PheWAS also apply to temporal-informed phenotyping. ICD codes are imperfect and may not fully describe the spectrum of symptoms or diagnoses experienced by patients, and furthermore ICD coding practices may vary among providers and across time.[84,85] Phecodes have differing levels of frequency in the EHR as well as varying levels of granularity. This may translate to varying positive and negative predictive values among different phecodes.[11,24] Our temporal-informed PheWAS also employed the standard statistical testing for PheWAS which uses logistic regression on binary phenotypes outcomes, but other statistical methods could have more favorable characteristics for detecting smaller associations. A simulation-based PheWAS study by Hughey et. al. reported that a time-to-event study design using Cox regression increased statistical power by approximately 10% when compared to standard PheWAS using binary logistic regression.[66] Finally, some limitations are specific to our temporal-based design. This study focused on identifying new diagnoses arising after a single discrete medical event, in this case COVID-19 testing. Some post-COVID phenotypes, particularly some symptoms, may be temporary and could have resolved by the end of the study observation period. Moreover, the approach treated time as two distinct periods, pre-recovery and post-recovery, without consideration of differences across time within the two temporal periods. Some diagnoses made in the remote past may not longer be significant or valid for individual patients, which may have increased the number of patients excluded for specific phenotypes and thus reduced statistical power for those analyses. Similarly, new diagnoses occurring soon after recovery may have a different likelihood of being related to a post-COVID syndrome compared to new diagnoses occurring many months later, so treating all diagnoses occurring within the post-recovery period with equal weight may not appropriately capture potential differences in relevance of some diagnoses to the post-COVID syndrome based on time of onset.

Future Directions of Research

Several of the limitations noted in the previous section merit further inquiry. In particular, further work addressing the potential time-varying nature of some phenotypes could be explored through time-to-event analyses or using time-based penalization of some pre-recovery phenotypes occurring in the remote past and post-recovery phenotypes occurring long after recovery. Replication in an external EHR dataset, particularly a public dataset with similar design features like the All of Us Research Program,[86,87] would further strengthen the generalizability of our results. Temporal-informed phenotyping could be readily adapted to assess the post-acute phenome among survivors of other serious acute medical events such as hospitalization for pneumonia, sepsis, or emergency general surgery, as these conditions are also associated with significant symptom burden and high rates of functional impairment in survivors.[79,80,88] Analyses that incorporate phenotypes extracted from alternative EHR data sources including procedural codes, clinical concepts extracted from narrative text as described by Zhao et. al.,[17] or comprehensive assessments of clinical laboratory studies as described by Dennis et. al.[15] could reveal additional disease patterns occurring among COVID-19 survivors. Finally, noting the original purpose of PheWAS was to discover genome-phenome associations, findings from temporal-informed PheWAS could be used to assess how the genetic architecture of new post-COVID diagnoses like atrial fibrillation, renal disease, pulmonary fibrosis, weight gain, etc. differs or conforms to the known genetic architecture of these diagnoses among patients without COVID-19. Potential datasets for assessing such post-COVID genome-phenome associations could include biobank data collected at our institution through the BioVU program,[89] as well as genetic datasets collected by national biobanks like All of Us[86,87] or multi-institutional genomics consortia such as the eMERGE Network or the COVID-19 Host Genetics Initiative.[90,91]

Conclusion

Temporal-informed phenotyping leverages a well-established high-throughput informatics phenotyping approach to provide a novel framework for understanding temporal changes in patients' longitudinal health status using EHR data. Initial evaluation of temporal-informed phenotyping demonstrates it readily identifies new diagnoses associated with COVID-19 survivorship using EHR data from VUMC. These findings replicate known associations in the medical literature and are supported by changes in several associated vital signs and laboratory tests. With the continued accumulation of EHR data across healthcare systems, temporal-informed phenotyping will enable future efforts to study the medical burdens experienced by survivors of acute medical illnesses.

REFERENCES

1 Adler-Milstein J, DesRoches CM, Kralovec P, *et al.* Electronic Health Record Adoption In US Hospitals: Progress Continues, But Challenges Persist. *Health Affairs* 2015;**34**:2174–80. doi:10.1377/hlthaff.2015.0992

2 Adler-Milstein J, Jha AK. HITECH Act Drove Large Gains In Hospital Electronic Health Record Adoption. *Health Affairs* 2017;**36**:1416–22. doi:10.1377/hlthaff.2016.1651

3 National Trends in Hospital and Physician Adoption of Electronic Health Records | HealthIT.gov. https://www.healthit.gov/data/quickstats/national-trends-hospital-and-physician-adoption-electronic-health-records (accessed 7 Sep 2022).

4 Danciu I, Cowan JD, Basford M, *et al.* Secondary Use of Clinical Data: the Vanderbilt Approach. *J Biomed Inform* 2014;**52**:28–35. doi:10.1016/j.jbi.2014.02.003

5 Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med* 2015;**7**:41. doi:10.1186/s13073-015-0166-y

6 Toga AW, Foster I, Kesselman C, *et al.* Big biomedical data as the key resource for discovery science. *J Am Med Inform Assoc* 2015;**22**:1126–31. doi:10.1093/jamia/ocv077

7 Hripcsak G, Ryan PB, Duke JD, *et al.* Characterizing treatment pathways at scale using the OHDSI network. *PNAS* 2016;**113**:7329–36. doi:10.1073/pnas.1510502113

8 Casey JA, Schwartz BS, Stewart WF, *et al.* Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health* 2016;**37**:61–81. doi:10.1146/annurev-publhealth-032315-021353

9 Jungkunz M, Köngeter A, Mehlis K, *et al.* Secondary Use of Clinical Data in Data-Gathering, Non-Interventional Research or Learning Activities: Definition, Types, and a Framework for Risk Assessment. *Journal of Medical Internet Research* 2021;**23**:e26631. doi:10.2196/26631

10 Richesson RL, Marsolo KS, Douthit BJ, *et al.* Enhancing the use of EHR systems for pragmatic embedded research: lessons from the NIH Health Care Systems Research Collaboratory. *J Am Med Inform Assoc* 2021;**28**:2626–40. doi:10.1093/jamia/ocab202

11 Denny JC, Bastarache L, Roden DM. Phenome-Wide Association Studies as a Tool to Advance Precision Medicine. *Annu Rev Genomics Hum Genet* 2016;**17**:353–73. doi:10.1146/annurev-genom-090314-024956

12 Bastarache L. Using Phecodes for Research with the Electronic Health Record: From PheWAS to PheRS. *Annu Rev Biomed Data Sci* 2021;**4**:1–19. doi:10.1146/annurev-biodatasci-122320-112352

13 Bastarache L, Hughey JJ, Hebbring S, *et al.* Phenotype risk scores identify patients with unrecognized Mendelian disease patterns. *Science* 2018;**359**:1233–9. doi:10.1126/science.aal4043

14 Ryan PB, Madigan D, Stang PE, *et al.* Medication-wide association studies. *CPT Pharmacometrics Syst Pharmacol* 2013;**2**:e76. doi:10.1038/psp.2013.52

15 Dennis JK, Sealock JM, Straub P, *et al.* Lab-wide association scan of polygenic scores identifies biomarkers of complex disease. *medRxiv* 2020;:2020.01.24.20018713. doi:10.1101/2020.01.24.20018713

16 Zheng NS, Feng Q, Kerchberger VE, *et al.* PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. *Journal of the American Medical Informatics Association* 2020;**27**:1675–87. doi:10.1093/jamia/ocaa104

17 Zhao J, Grabowska ME, Kerchberger VE, *et al.* ConceptWAS: A high-throughput method for early identification of COVID-19 presenting symptoms and characteristics from clinical notes. *J Biomed Inform* 2021;**117**:103748. doi:10.1016/j.jbi.2021.103748

18 Wu P, Nelson SD, Zhao J, *et al.* DDIWAS: High-throughput electronic health record-based screening of drug-drug interactions. *J Am Med Inform Assoc* 2021;**28**:1421–30. doi:10.1093/jamia/ocab019

19 Denny JC, Ritchie MD, Basford MA, *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* 2010;**26**:1205–10. doi:10.1093/bioinformatics/btq126

20 Pendergrass SA, Brown-Gentry K, Dudek SM, *et al.* The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genet Epidemiol* 2011;**35**:410–22. doi:10.1002/gepi.20589

21 Pendergrass SA, Brown-Gentry K, Dudek S, *et al.* Phenome-Wide Association Study (PheWAS) for Detection of Pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* 2013;**9**. doi:10.1371/journal.pgen.1003087

22 Wu P, Gifford A, Meng X, *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* 2019;**7**:e14325. doi:10.2196/14325

23 Denny JC, Bastarache L, Ritchie MD, *et al.* Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat Biotechnol* 2013;**31**:1102–10. doi:10.1038/nbt.2749

24 Wei W-Q, Bastarache LA, Carroll RJ, *et al.* Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLOS ONE* 2017;**12**:e0175508. doi:10.1371/journal.pone.0175508

25 Warner JL, Zollanvari A, Ding Q, *et al.* Temporal phenome analysis of a large electronic health record cohort enables identification of hospital-acquired complications. *J Am Med Inform Assoc* 2013;**20**:e281–7. doi:10.1136/amiajnl-2013-001861

26 Verma A, Leader JB, Verma SS, *et al.* INTEGRATING CLINICAL LABORATORY MEASURES AND ICD-9 CODE DIAGNOSES IN PHENOME-WIDE ASSOCIATION STUDIES. *Pac Symp Biocomput* 2016;**21**:168–79.

27 Cai W, Cagan A, He Z, *et al.* A Phenome-Wide Analysis of Healthcare Costs Associated with Inflammatory Bowel Diseases. *Dig Dis Sci* 2021;**66**:760–7. doi:10.1007/s10620-020-06329-9

28 Dashti HS, Cade BE, Stutaite G, *et al.* Sleep health, diseases, and pain syndromes: findings from an electronic health record biobank. *Sleep* 2020;**44**:zsaa189. doi:10.1093/sleep/zsaa189

29 Zhang T, Goodman M, Zhu F, *et al.* Phenome-wide examination of comorbidity burden and multiple sclerosis disease severity. *Neurol Neuroimmunol Neuroinflamm* 2020;**7**. doi:10.1212/NXI.0000000000000864

30 Pulley JM, Jerome RN, Bernard GR, *et al.* The Astounding Breadth of Health Disparity: Phenome-Wide Effects of Race on Disease Risk. *J Natl Med Assoc* 2021;**113**:187–94. doi:10.1016/j.jnma.2020.08.009

31 Niarchou M, Lin GT, Lense MD, *et al.* Medical phenome of musicians: an investigation of health records collected on 9803 musically active individuals. *Ann N Y Acad Sci* 2021;**1505**:156–68. doi:10.1111/nyas.14671

32 Carroll RJ, Bastarache L, Denny JC. R PheWAS: data analysis and plotting tools for phenome-wide association studies in the R environment. *Bioinformatics* 2014;**30**:2375–6. doi:10.1093/bioinformatics/btu197

33 Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020;**20**:533–4. doi:10.1016/S1473-3099(20)30120-1

34 Irons NJ, Raftery AE. Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. *Proceedings of the National Academy of Sciences* 2021;**118**:e2103272118. doi:10.1073/pnas.2103272118

35 CDC. Estimated COVID-19 Burden. Centers for Disease Control and Prevention. 2020.https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/burden.html (accessed 7 Sep 2022).

36 BMJ. Why we need to keep using the patient made term “Long Covid.” The BMJ. 2020.https://blogs.bmj.com/bmj/2020/10/01/why-we-need-to-keep-using-the-patient-made-term-long-covid/ (accessed 12 Sep 2022).

37 Callard F, Perego E. How and why patients made Long Covid. *Social Science & Medicine* 2021;**268**:113426. doi:10.1016/j.socscimed.2020.113426

38 Nalbandian A, Sehgal K, Gupta A, *et al.* Post-acute COVID-19 syndrome. *Nat Med* 2021;**27**:601–15. doi:10.1038/s41591-021-01283-z

39 Datta SD, Talwar A, Lee JT. A Proposed Framework and Timeline of the Spectrum of Disease Due to SARS-CoV-2 Infection: Illness Beyond Acute Infection and Public Health Implications. *JAMA* 2020;**324**:2251. doi:10.1001/jama.2020.22717

40 Logue JK, Franko NM, McCulloch DJ, *et al.* Sequelae in Adults at 6 Months After COVID-19 Infection. *JAMA Netw Open* 2021;**4**:e210830. doi:10.1001/jamanetworkopen.2021.0830

41 Ayoubkhani D, Khunti K, Nafilyan V, *et al.* Post-covid syndrome in individuals admitted to hospital with covid-19: retrospective cohort study. *BMJ* 2021;**372**:n693. doi:10.1136/bmj.n693

42 Morin L, Laurent Savale, Pham T, *et al.* Four-Month Clinical Status of a Cohort of Patients After Hospitalization for COVID-19. *JAMA* 2021;**325**:1525–34. doi:10.1001/jama.2021.3331

43 Sonnweber T, Sahanic S, Pizzini A, *et al.* Cardiopulmonary recovery after COVID-19: an observational prospective multicentre trial. *Eur Respir J* 2021;**57**. doi:10.1183/13993003.03481-2020

44 Arnold DT, Hamilton FW, Milne A, *et al.* Patient outcomes after hospitalisation with COVID-19 and implications for follow-up: results from a prospective UK cohort. *Thorax* 2021;**76**:399–401. doi:10.1136/thoraxjnl-2020-216086

45 Daugherty SE, Guo Y, Heath K, *et al.* Risk of clinical sequelae after the acute phase of SARS-CoV-2 infection: retrospective cohort study. *BMJ* 2021;**373**:n1098. doi:10.1136/bmj.n1098

46 Blanco J-R, Cobos-Ceballos M-J, Navarro F, *et al.* Pulmonary long-term consequences of COVID-19 infections after hospital discharge. *Clin Microbiol Infect* 2021;**27**:892–6. doi:10.1016/j.cmi.2021.02.019

47 Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. *Nature* 2021;**594**:259–64. doi:10.1038/s41586-021-03553-9

48 Taquet M, Geddes JR, Husain M, *et al.* 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. doi:10.1016/S2215-0366(21)00084-5

49 Davis HE, Assaf GS, McCorkell L, *et al.* Characterizing long COVID in an international cohort: 7 months of symptoms and their impact. *EClinicalMedicine* 2021;**38**. doi:10.1016/j.eclinm.2021.101019

50 Huang L, Yao Q, Gu X, *et al.* 1-year outcomes in hospital survivors with COVID-19: a longitudinal cohort study. *The Lancet* 2021;**398**:747–58. doi:10.1016/S0140-6736(21)01755-4

51 Estiri H, Strasser ZH, Brat GA, *et al.* Evolving phenotypes of non-hospitalized patients that indicate long COVID. *BMC Medicine* 2021;**19**:249. doi:10.1186/s12916-021-02115-0

52 Bull-Otterson L. Post–COVID Conditions Among Adult COVID-19 Survivors Aged 18–64 and ≥65 Years — United States, March 2020–November 2021. *MMWR Morb Mortal Wkly Rep* 2022;**71**. doi:10.15585/mmwr.mm7121e1

53 Soriano JB, Murthy S, Marshall JC, *et al.* A clinical case definition of post-COVID-19 condition by a Delphi consensus. *The Lancet Infectious Diseases* 2022;**22**:e102–7. doi:10.1016/S1473-3099(21)00703-9

54 National Institute for Health Care and Excellent. COVID-19 rapid guideline: managing the long-term effects of COVID-19. https://www.nice.org.uk/guidance/ng188 (accessed 13 Sep 2022).

55 Centers for Disease Control and Prevention. Post-COVID Conditions. Centers for Disease Control and Prevention. 2022.https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html (accessed 13 Sep 2022).

56 Carfì A, Bernabei R, Landi F, *et al.* Persistent Symptoms in Patients After Acute COVID-19. *JAMA* 2020;**324**:603. doi:10.1001/jama.2020.12603

57 Deer RR, Rock MA, Vasilevsky N, *et al.* Characterizing Long COVID: Deep Phenotype of a Complex Condition. *eBioMedicine* 2021;**74**. doi:10.1016/j.ebiom.2021.103722

58 Huang L, Li X, Gu X, *et al.* Health outcomes in people 2 years after surviving hospitalisation with COVID-19: a longitudinal cohort study. *The Lancet Respiratory Medicine* 2022;**10**:863–76. doi:10.1016/S2213-2600(22)00126-6

59 Tenforde MW, Kim SS, Lindsell CJ, *et al.* Symptom Duration and Risk Factors for Delayed Return to Usual Health Among Outpatients with COVID-19 in a Multistate Health Care Systems Network - United States, March-June 2020. *MMWR Morb Mortal Wkly Rep* 2020;**69**:993–8. doi:10.15585/mmwr.mm6930e1

60 Havervall S, Rosell A, Phillipson M, *et al.* Symptoms and Functional Impairment Assessed 8 Months After Mild COVID-19 Among Health Care Workers. *JAMA* 2021;**325**:2015–6. doi:10.1001/jama.2021.5612

61 Nehme M, Braillard O, Chappuis F, *et al.* Prevalence of Symptoms More Than Seven Months After Diagnosis of Symptomatic COVID-19 in an Outpatient Setting. *Ann Intern Med* 2021;**174**:1252–60. doi:10.7326/M21-0878

62 Blomberg B, Mohn KG-I, Brokstad KA, *et al.* Long COVID in a prospective cohort of home-isolated patients. *Nat Med* 2021;**27**:1607–13. doi:10.1038/s41591-021-01433-3

63 Sneller MC, Liang CJ, Marques AR, *et al.* A Longitudinal Study of COVID-19 Sequelae and Immunity: Baseline Findings. *Ann Intern Med* 2022;**175**:969–79. doi:10.7326/M21-4905

64 Oetjens MT, Luo JZ, Chang A, *et al.* Electronic health record analysis identifies kidney disease as the leading risk factor for hospitalization in confirmed COVID-19 patients. *PLOS ONE* 2020;**15**:e0242182. doi:10.1371/journal.pone.0242182

65 Salvatore M, Gu T, Mack JA, *et al.* A Phenome-Wide Association Study (PheWAS) of COVID-19 Outcomes by Race Using the Electronic Health Records Data in Michigan Medicine. *J Clin Med* 2021;**10**:1351. doi:10.3390/jcm10071351

66 Hughey JJ, Rhoades SD, Fu DY, *et al.* Cox regression increases power to detect genotype-phenotype associations in genomic studies using the electronic health record. *BMC Genomics* 2019;**20**:805. doi:10.1186/s12864-019-6192-1

67 DeLozier S, Bland S, McPheeters M, *et al.* Phenotyping coronavirus disease 2019 during a global health pandemic: Lessons learned from the characterization of an early cohort. *J Biomed Inform* 2021;**117**:103777. doi:10.1016/j.jbi.2021.103777

68 Wang D, Hu B, Hu C, *et al.* Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. *JAMA* 2020;**323**:1061. doi:10.1001/jama.2020.1585

69 Yang X, Yu Y, Xu J, *et al.* Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: a single-centered, retrospective, observational study. *Lancet Respir Med* 2020;**8**:475–81. doi:10.1016/S2213-2600(20)30079-5

70 Huang C, Huang L, Wang Y, *et al.* 6-month consequences of COVID-19 in patients discharged from hospital: a cohort study. *The Lancet* 2021;**397**:220–32. doi:10.1016/S0140-6736(20)32656-8

71 Feng Q, Wei W-Q, Chaugai S, *et al.* Association Between Low-Density Lipoprotein Cholesterol Levels and Risk for Sepsis Among Patients Admitted to the Hospital With Infection. *JAMA Netw Open* 2019;**2**:e187223. doi:10.1001/jamanetworkopen.2018.7223

72 Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLOS Medicine* 2015;**12**:e1001885. doi:10.1371/journal.pmed.1001885

73 Wang SV, Pinheiro S, Hua W, *et al.* STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *BMJ* 2021;**372**:m4856. doi:10.1136/bmj.m4856

74 Zhang D, Shen X, Qi X. Resting heart rate and all-cause and cardiovascular mortality in the general population: a meta-analysis. *CMAJ* 2016;**188**:E53–63. doi:10.1503/cmaj.150535

75 Meng W, Ou W, Chandwani S, *et al.* Temporal phenotyping by mining healthcare data to derive lines of therapy for cancer. *J Biomed Inform* 2019;**100**:103335. doi:10.1016/j.jbi.2019.103335

76 Zhao J, Zhang Y, Schlueter DJ, *et al.* Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: Cardiovascular disease case study. *J Biomed Inform* 2019;**98**:103270. doi:10.1016/j.jbi.2019.103270

77 Kim Y, Lhatoo S, Zhang G-Q, *et al.* Temporal Phenotyping for Transitional Disease Progress: an Application to Epilepsy and Alzheimer's Disease. *J Biomed Inform* 2020;**107**:103462. doi:10.1016/j.jbi.2020.103462

78 Pfaff ER, Girvin AT, Bennett TD, *et al.* Identifying who has long COVID in the USA: a machine learning approach using N3C data. *The Lancet Digital Health* 2022;**4**:e532–41. doi:10.1016/S2589-7500(22)00048-6

79 Yende S, Linde-Zwirble W, Mayr F, *et al.* Risk of Cardiovascular Events in Survivors of Severe Sepsis. *Am J Respir Crit Care Med* 2014;**189**:1065–74. doi:10.1164/rccm.201307-1321OC

80 Corrales-Medina VF, Alvarez KN, Weissfeld LA, *et al.* Association Between Hospitalization for Pneumonia and Subsequent Risk of Cardiovascular Disease. *JAMA* 2015;**313**:264–74. doi:10.1001/jama.2014.18229

81 Haendel MA, Chute CG, Bennett TD, *et al.* The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* Published Online First: 17 August 2020. doi:10.1093/jamia/ocaa196

82 Bhatt AS, McElrath EE, Claggett BL, *et al.* Accuracy of ICD-10 Diagnostic Codes to Identify COVID-19 Among Hospitalized Patients. *J Gen Intern Med* 2021;**36**:2532–5. doi:10.1007/s11606-021-06936-w

83 Hodcroft EB. CoVariants: SARS-CoV-2 Mutations and Variants of Interest. 2021.https://covariants.org/ (accessed 18 Aug 2021).

84 O'Malley KJ, Cook KF, Price MD, *et al.* Measuring Diagnoses: ICD Code Accuracy. *Health Serv Res* 2005;**40**:1620–39. doi:10.1111/j.1475-6773.2005.00444.x

85 Sivashankaran S, Borsi JP, Yoho A. Have ICD-10 Coding Practices Changed Since 2015? *AMIA Annu Symp Proc* 2020;**2019**:804–11.

86 The All of Us Research Program Investigators. The "All of Us" Research Program. *New England Journal of Medicine* 2019;**381**:668–76. doi:10.1056/NEJMsr1809937

87 Research Program. COVID-19 Research Initiatives – All of Us Research Hub. https://www.researchallofus.org/discover/covid-19-research-initiatives/ (accessed 15 Sep 2022).

88 Smith JW, Davis JK, Quatman-Yates CC, *et al.* Loss of Community-Dwelling Status Among Survivors of High-Acuity Emergency General Surgery Disease. *J Am Geriatr Soc* 2019;**67**:2289–97. doi:10.1111/jgs.16046

89 Roden D, Pulley J, Basford M, *et al.* Development of a Large-Scale De-Identified DNA Biobank to Enable Personalized Medicine. *Clin Pharmacol Ther* 2008;**84**:362–9. doi:10.1038/clpt.2008.89

90 Kho AN, Pacheco JA, Peissig PL, *et al.* Electronic medical records for genetic research: results of the eMERGE consortium. *Science Translational Medicine* 2011;**3**:79re1. doi:10.1126/scitranslmed.3001807

91 Niemi MEK, Karjalainen J, Liao RG, *et al.* Mapping the human genetic architecture of COVID-19. *Nature* 2021;**600**:472–7. doi:10.1038/s41586-021-03767-x

92 FitzHenry F, Resnic FS, Robbins SL, *et al.* Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. *Appl Clin Inform* 2015;**6**:536–47. doi:10.4338/ACI-2014-12-CR-0121

93 Denny JC, Giuse DA, Jirjis JN. The Vanderbilt Experience with Electronic Health Records. *Semin Colon Rectal Surg* 2005;**16**:59–68. doi:10.1053/j.scrs.2005.08.003

94 Quan H, Sundararajan V, Halfon P, *et al.* Coding Algorithms for Defining Comorbidities in ICD-9-CM and ICD-10 Administrative Data. *Medical Care* 2005;**43**:1130–9. doi:10.1097/01.mlr.0000182534.19832.83

95 Ho D, Imai K, King G, *et al.* MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software* 2011;**42**:1–28. doi:10.18637/jss.v042.i08

96 Greifer N. cobalt: Covariate Balance Tables and Plots. 2022.https://ngreifer.github.io/cobalt/

APPENDIX A

VUMC EHR database and data cleaning

Study setting

Study participants came from Vanderbilt University Medical Center (VUMC), a private nonprofit academic medical institution based in Nashville, Tennessee, USA. It is one of the largest medical centers in the southeastern United States and serves as an anchor for specialty and primary care for patients throughout Tennessee and the Mid-South region. VUMC has maintained an EHR since the mid-1990's and largely eliminated paper records in clinical care since 2004.[4] In 2019 before the start of the pandemic, VUMC had 1,131 licensed beds, and managed over 2 million ambulatory visits, over 110,000 emergency department visits, and over 55,000 surgical procedures annually.

Vanderbilt Research Derivative (RD)

The Research Derivative is a database of clinical and related data derived from the Vanderbilt University Medical Center's (VUMC) clinical systems and restructured for research.[4] Data is repurposed from VUMC's enterprise data warehouse (EDW), which includes data from clinical information systems eStar (VUMC's local implementation of Epic Hyperspace®), VPIMS (Vanderbilt Perioperative Information Management Systems), ORMIS (Operating Room Management Information System), and clinical laboratory systems, as well as legacy systems including StarPanel (Vanderbilt's native electronic medical records system), Horizon Export Orders, and others. Data is transformed from the EDW to the RD using a set of custom Extract, Transform, and Load (ETL) pipelines to map data to the Observational Medical Outcomes Partnership (OMOP) common data model.[92] The medical record number and other person identifiers are preserved within the database. Data types include reimbursement codes, clinical notes and documentation, nursing records, medication data, laboratory data, encounter and visit data, and respiratory flowsheet data on mechanical ventilator and oxygen usage. Output may include structured data points, such as International Classification of Diseases (ICD) codes and encounter dates, semi-structured data such as laboratory tests and results, or unstructured data such as physician progress reports. Pertinent to this study, laboratory tests are coded as Logical Observation Identifiers Names and Codes (LOINC) and diagnoses are coded using the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM). The RD database is stored on a secure database server housed in the Vanderbilt Data Center. The database is fully compliant with the administrative, physical, and technical provisions of the Health Insurance Portability and

Accountability Act of 1996 (HIPAA) Security and Privacy Rules, and operates with oversight from the Vanderbilt Institutional Review Board. The database is maintained by VUMC's Office of Research Informatics and the Vanderbilt Institute for Clinical and Translational Research (VICTR) Big Data team under the direction of Paul Harris, PhD.

Vanderbilt COVID-19 registry database

The COVID-19 registry utilizes a customized version of the RD specifically created to support COVID-related research efforts.[67] It captures all patients who had SARS-CoV-2 testing performed at VUMC and its associated community testing sites, outpatient clinics, urgent care centers, and affiliate hospitals. The registry collects a broad range of health data including demographics, vital signs, medical and social history, medications, self-reported symptoms, visit information, clinic notes, respiratory flowsheets, diagnostic codes, and procedural codes. The COVID-19 registry is refreshed daily to facilitate close to real-time access to research data, rather than the usual monthly refresh for the "traditional" RD as described above. Additional COVID-19-specific tables are included in the registry database specifically related to SARS-CoV-2 testing including testing dates, care sites, ordering provider, test status, indication for testing (symptoms consistent with clinical COVID-19, or one of several institutionally-determined acceptable indications for asymptomatic testing), along with test results.

Investigator access to database population

The study authors had full access to all patient and clinical data available in the RD and COVID-19 registry database.

Quality control of study data

Patient-level quality control. We excluded nine patients with missing data on date of birth or sex, or had zero visit records present in the EHR (ire. even no visit associated with a SARS-CoV-2 PCR test). We included patients who had race or ethnicity coded as "Unknown / Not reported" as some patients choose not to provide a self-identified race or ethnicity when being registered for a care visit.

SARS-CoV-2 testing data quality control. we identified the records of all SARS-CoV-2 PCR tests included in VUMC's research derivative using the LOINC code 94533-7. We excluded all tests that were performed after the data censoring date (January 1, 2022) and those with invalid dates or clearly erroneous dates (e.g. testing dates before February 1, 2020). We only included test results which were identified as completed, and not still in progress or never resulted. We also only included those laboratory test records which had a result of either “Positive” or “Not Detected”, and excluded any PCR tests with indeterminate, pending, or canceled results.

ICD code and phecode data quality control. As the COVID-19 registry database is updated daily, some data is subject to change as they are finalized, amended, corrected, or updated in the medical record. Pertinent to this study, ICD diagnosis codes are typically entered into the enterprise data warehouse by practitioners, but can be changed during finalization by coding specialists several days or weeks later. Our experience and that of the VICTR Big Data staff has been that ICD codes and problem lists have relatively low volatility over time.[4] To mitigate the potential for any changes in diagnosis codes, we extracted the diagnosis codes from the registry database in late February 2022 which was several weeks after the study’s data censoring date.

Vital sign and laboratory value data quality control. We captured vital signs and clinical laboratory test results (Appendix D) for each patient obtained during outpatient encounters for two separate time period: (1) results within 180 days before the index SARS-CoV-2 and (2) results within 365 days after the recovery date. We took the median value for each period for final analyses. Similar to our experience with diagnosis codes, we have found relatively low volatility for these results once they are finalized in the EHR.[4]

APPENDIX B

PheWAS model design, covariates, and sensitivity analyses

Exclusions from specific phenotype analyses.

As described in the **Methods** section of the main manuscript, for each phenotype we excluded any patient who had the corresponding phecode in their medical record prior to reaching the recovery period. Reasons for this included (1) diagnoses entered into the EHR prior to the index SARS-CoV-2 PCR test, (2) diagnoses made during an inpatient hospitalization associated with the index PCR test, or (3) diagnoses made before the patient reached the 30-day recovery period. We allowed an unlimited look-back period to identify phecodes entered prior to SARS-CoV-2 testing to minimize the risk of misclassifying old diagnoses as "new" in the post-acute period. The electronic health record at Vanderbilt dates back to the early 1990s,[93] but in practice the overwhelming majority of pre-testing phecodes (98.9%) in our cohort were for diagnosis occurring within the last 20 years (after 2001).

Phecode analyses

For PheWAS analyses using phecodes, we required a minimum of at least 10 phenotype cases in the overall cohort. Covariates included in the PheWAS models included age, sex, race (white, black, other/multi-ethnic, or not reported), ethnicity (Hispanic/Latino, Non-Hispanic/Non-Latino, or not reported), time under observation in the post-acute period (days), testing indication (symptomatic or asymptomatic testing), and Charlson index comorbidities present prior to SARS-CoV-2 testing using a phecode-based definition (***PheWAS Model Variables*** below; Appendix C).[71,94]

Vital sign / clinical laboratory test analyses

Covariates included in linear regression models evaluating changes in vital sign or clinical laboratory tests included age, sex, race, ethnicity, and time between the pre-testing value and the post-acute phase value (in days).

Loss of follow-up and missing phecode data

All patients who underwent SARS-CoV-2 testing were included in the primary analysis regardless whether they had a subsequent visit at VUMC. This is in keeping with usual practice for

PheWAS as the methodology assumes that for each phenotype patients without relevant diagnosis code did not develop the phenotype and thus can be considered controls.[19]

Sensitivity analyses

We assessed the robustness of our findings to our model assumptions using several sensitivity analyses. Firstly, we assessed the effect of our control definition and loss to follow-up by limiting the analysis to those patients who had at least one follow-up visit during the post-acute phase. Secondly, we assessed the effect of patients newly seeking care in our system during the pandemic by evaluating only those patients who had at least 2 (two) visits in our system separated by at least 6 months prior to the start of the pandemic. Thirdly, we assessed the effects of the standard PheWAS "phenotype case" by using a more relaxed "phenotype case" definition requiring a phecode only on a single visit. Lastly, we assessed the effect of bias from differences in baseline clinical variables using a propensity-matched sub-cohort which matched 3 never-infected controls to each COVID-19 survivor. Further details are provided below.

PheWAS model variables

1) **Age** at time of testing (years)
2) **Sex** coded as female or male.
3) **Race** coded as 4 separate binary variables: White, Black or African American, Other race or multi-racial, or Unknown/not reported.
4) **Ethnicity** coded as 3 separate binary variables: Hispanic or Latino, Non-Hispanic/Non-Latino, or Unknown/not reported.
5) **Post-acute observation time**. Number of days from start of post-acute phase (as defined in **Methods** of the main manuscript) until date of data censoring.
6) **Asymptomatic testing**. Binary variable indicating the patient's SARS-CoV-2 test was performed for asymptomatic screening versus symptomatic testing. Reasons for asymptomatic screening included asymptomatic admission to the hospital for another diagnosis, pre-procedural or pre-surgical screening, known SARS-CoV-2 exposure, pre-receipt of immunosupressive or anti-neoplastic therapy, pre-transplant evaluation, or requirement for placement in post-acute care or long-term nursing care.
7) **Inpatient hospitalization**. Binary variable indicating if the SARS-CoV-2 test was performed within 15 days prior to an inpatient hospitalization, or performed during an inpatient hospitalization.
8) **Charlson comorbidities**. Each comorbidity is encoded as a binary variable indicating presence of corresponding phecode(s) on one or more visits at least 15 days prior to SARS-CoV-2 test (See **Table S2**).
   i. Myocardial infarction.
   ii. Congestive heart failure
   iii. Peripheral vascular disease
   iv. Cerebrovascular disease

v. Dementia
vi. Chronic pulmonary disease
vii. Rheumatologic disease
viii. Peptic ulcer disease
ix. Diabetes
x. Mild liver disease
xi. Severe liver disease
xii. Hemiplegia or paraplegia
xiii. Renal disease
xiv. Any malignancy, including lymphoma or leukemia, not non-melanoma skin cancer.
xv. Metastatic solid tumor
xvi. AIDS or HIV infection

Propensity matching analysis

As imbalances in some clinical variables were noted between the SARS-CoV-2 positive and never-infected groups, we compared the results of performing a PheWAS using the full (unmatched) study cohort with results using a propensity-score matched sub-cohort. We generated a propensity score to estimate probability of having a positive SARS-CoV-2 test using a generalized linear model with a probit link function. We conditioned the propensity scoring model on age, sex, race, ethnicity, symptomatic testing indication, inpatient hospitalization around time of SARS-CoV-2 test, observation time after recovery (in days), and length of EHR prior to SARS-CoV-2 testing (in years). We then performed nearest neighbor matching without replacement with a 3:1 control-case ratio. Control-case ratios of 1:1 and 2:1 were also assessed but did not result in satisfactory matching performance. After matching, all standardized mean differences in the conditioning variables were below 0.1 and Kolmogorov-Smirnov statistics (maximum difference in empirical cumulative density function) demonstrated improvement for all variables, indicating acceptable matching between cases and controls (Appendix H). Visual assessment of continuously distributed conditioning variables (age, observation time after recovery, and EHR length prior to SARS-CoV-2 testing) demonstrated that the matching improved alignment between the SARS-CoV-2 positive and never-infected control group for all variables. We did note a modest persistent difference in distributions of observation time after recovery between groups, with the SARS-CoV-2 group having a more peaked distribution of post-recovery observation time versus a more uniform distribution among the never-infected controls. This reflected the higher positive test rates during local waves of the pandemic, whereas the distribution of negative tests was spread more uniformly over time. Noting this modest persistent difference between exposure groups for post-recovery observation time, post-matching standardized mean difference for this variable remained acceptable at 0.045. Matching was performed using the R package *MatchIt* and visualization of covariate balance was performed using the R package *cobalt.[95,96]*

After matching, all SARS-CoV-2 cases (n=30,088) and the 90,264 never-infected controls were included in the propensity-matched analysis. We repeated the PheWAS analysis using the same covariates as in the primary analysis and compared number of significant phecode associations and effects odds ratios between the full unmatched cohort (primary analysis) and the propensity-matched cohort. Results of the propensity-matched PheWAS are reported in Appendix H.

APPENDIX C

Phecode groupings to identify Charlson comorbidities

Comorbidity	**Phecode**
Myocardial infarction	411.2
Congestive heart failure	428, 428.1, 428.2, 428.3, 428.4
Peripheral vascular disease	443.9, 440, 440.2, 440.21, 440.22, 442.1, 442.11, 791, 459.7
Cerebrovascular disease	430, 430.1, 430.2, 430.3, 433, 433.1, 433.11, 433.12, 433.2, 433.21, 433.3, 433.31, 433.32, 433.5, 433.6, 433.8
Dementia	290, 290.1, 290.13, 290.16
Chronic pulmonary disease	497, 496.2, 496.21, 496.1, 495, 495.2, 495.1, 495.11, 496.3, 500.1, 500.2, 496, 500
Rheumatologic disease	695.42, 709.2, 709.3, 709.4, 709.5, 709.6, 709.7, 714, 714.1, 714.2, 717
Peptic ulcer disease	531, 531.1, 531.2, 531.3, 531.4, 531.5
Diabetes	250, 250.1, 250.11, 250.12, 250.13, 250.14, 250.15, 250.2, 250.21, 250.22, 250.23, 250.24, 250.25, 250.3
Mild liver disease	317.11, 070, 070.1, 070.2, 070.3, 070.4, 571.5, 571.51, 571.6
Severe liver disease	530.2, 571.8, 571.81
Hemiplegia or paraplegia	334.1, 342, 343, 344
Renal disease	401.2, 401.22, 580.1, 580.11, 580.12, 580.14, 580.3, 580.31, 580.32, 585.4, 585.34, 585.2, 588, 588.1, 588.2
Any malignancy, including lymphoma or leukemia, excluding non-melanoma skin cancer.	145, 145.1, 145.2, 145.3, 145.4, 145.5, 149, 149.1, 149.2, 149.3, 149.4, 145.9, 150, 151, 153, 153.2, 153.3, 155, 155.1, 157, 158, 159, 159.2, 159.3, 159.4, 164, 165, 165.1, 170, 170.1, 170.2, 172.11, 174, 174.1, 174.11, 174.2, 174.3, 180, 180.1, 180.3, 182, 184, 184.1, 184.2, 185, 187, 187.1, 187.2, 187.8, 189, 189.1, 189.11, 189.12, 189.2, 189.21, 189.4, 190, 191, 191.1, 191.11, 193, 194, 195, 195.1, 195.3, 200, 200.1, 201, 202, 202.2, 202.21, 202.22, 202.23, 202.24, 204, 204.1, 204.11, 204.12, 204.2, 204.21, 204.22, 204.3, 204.4, 202.22, 202.23, 202.24
Metastatic solid tumor	195.1, 198, 198.1, 198.2, 198.3, 198.4, 198.5, 198.6, 198.7
AIDS or HIV infection	071, 071.1

APPENDIX D

Charlson comorbidities, vital signs, and laboratory data of study population

Characteristic	Never Infected	SARS-CoV-2 Positive	Overall
Number in cohort	156,017	30,088	186,105
Comorbidites prior to SARS-CoV-2 test (%) [a]			
Congestive heart failure	4,678 (3.0)	630 (2.1)	5,308 (2.9)
Diabetes	10,702 (6.9)	1,964 (6.5)	12,666 (6.8)
Myocardial infarction	2,880 (1.8)	419 (1.4)	3,299 (1.8)
Peripheral vascular disease	2,464 (1.6)	325 (1.1)	2,789 (1.5)
Cerebrovascular disease	4,289 (2.7)	610 (2.0)	4,899 (2.6)
Dementia	787 (0.5)	137 (0.5)	924 (0.5)
Chronic pulmonary disease	9,580 (6.1)	1,646 (5.5)	11,226 (6.0)
Rheumatologic disease	3,097 (2.0)	428 (1.4)	3,525 (1.9)
Peptic ulcer disease	876 (0.6)	147 (0.5)	1,023 (0.5)
Mild liver disease	4,941 (3.2)	726 (2.4)	5,667 (3.0)
Severe liver disease	1,448 (0.9)	170 (0.6)	1,618 (0.9)
Hemiplegia or paraplegia	1,035 (0.7)	153 (0.5)	1,188 (0.6)
Renal disease	3,978 (2.5)	750 (2.5)	4,728 (2.5)
Any malignancy	12,045 (7.7)	1,463 (4.9)	13,508 (7.3)
Metastatic solid tumor	2,129 (1.4)	246 (0.8)	2,375 (1.3)
AIDS or HIV infection	1,065 (0.7)	173 (0.6)	1,238 (0.7)
Vital sign or laboratory values prior to test, (mean, SD)			
Body Mass Index (kg/m^2)	24.81 (3.12)	24.81 (3.07)	24.81 (3.11)
Systolic Blood Pressure (mmHg)	121 (15)	120 (13)	121 (14)
Heart Rate (bpm)	77 (11)	77 (11)	77 (11)
Respiratory Rate (min^{-1})	17 (2)	17 (2)	17 (2)
Oxygen saturation by pulse oximetry (SpO2, %)	98 (2)	98 (2)	98 (2)
White Blood Cell Count (10^3/µL)	6.8 (1.6)	6.8 (1.6)	6.8 (1.6)
Hemoglobin (gm/dL)	13.8 (1.5)	14.0 (1.4)	13.9 (1.4)
Platelet Count (10^3/µL)	247 (55)	250 (53)	247 (54)
Serum Potassium (mEq/L)	4.1 (0.3)	4.1 (0.3)	4.1 (0.3)
Serum Creatinine (mg/dL)	0.92 (0.21)	0.92 (0.19)	0.92 (0.20)
Estimated glomerular filtration rate (ml/min)	82 (19)	83 (18)	82 (19)
Hemoglobin A1C (%)	5.6 (0.7)	5.6 (0.6)	5.6 (0.7)
Serum glucose (mg/dL)	98 (20)	98 (19)	98 (20)
Time between pre-testing and post-recovery vital sign or laboratory assessment, in days. (mean, SD).[b]			
BMI	206 (123)	214 (113)	207 (122)
Pulse	219 (124)	250 (115)	224 (123)
Blood pressure	225 (123)	249 (114)	228 (122)
Respiratory rate	244 (126)	252 (113)	246 (124)
Oxygen saturation by pulse oximetry	242 (127)	252 (112)	244 (125)
White blood cell count	268 (129)	263 (112)	267 (127)
Hemoglobin level	263 (124)	274 (113)	265 (122)
Platelet count	265 (128)	263 (112)	265 (125)
Serum potassium	258 (124)	269 (112)	260 (122)
Estimated glomerular filtration rate	259 (125)	268 (112)	261 (123)
Hemoglobin A1C	255 (125)	265 (113)	257 (123)
Serum glucose	319 (108)	314 (101)	318 (107)

[a] Phecodes used to identify Charlson comorbidities are provided in Supplementary Table S2.

[b] Pre-testing values were collected up to 180 days prior to index SARS-CoV-2 testing for all vitals and lab values, and post-recovery values were allow up to 1 year after entering the post-recovery period.

APPENDIX E

Comparison of case retention across select phecode chapters.

Phecode chapter	Case retention (Median [IQR])	Comparator chapter	Case retention (Median [IQR])	p
neoplasms	18% [11% - 33%]	infectious diseases	40% [24% - 56%]	≤ 0.0001 *
’’	’’	endocrine/metabolic	28% [19% - 44%]	≤ 0.0001 *
’’	’’	hematopoietic	35% [24% - 45%]	0.00024 *
’’	’’	mental disorders	35% [24% - 56%]	≤ 0.0001 *
’’	’’	neurological	35% [22% - 48%]	≤ 0.0001 *
’’	’’	sense organs	38% [27% - 50%]	≤ 0.0001 *
’’	’’	circulatory system	32% [24% - 43%]	≤ 0.0001 *
’’	’’	respiratory	33% [26% - 51%]	≤ 0.0001 *
’’	’’	digestive	32% [16% - 50%]	≤ 0.0001 *
’’	’’	genitourinary	41% [30% - 57%]	≤ 0.0001 *
’’	’’	pregnancy complications	47% [37% - 60%]	≤ 0.0001 *
’’	’’	dermatologic	48% [28% - 59%]	≤ 0.0001 *
’’	’’	musculoskeletal	47% [33% - 58%]	≤ 0.0001 *
’’	’’	congenital anomalies	25% [18% - 43%]	0.0016
’’	’’	symptoms	46% [38% - 57%]	≤ 0.0001 *
’’	’’	injuries & poisonings	42% [32% - 55%]	≤ 0.0001 *
dermatologic	48% [28% - 59%]	infectious diseases	40% [24% - 56%]	0.32
"	"	neoplasms	18% [11% - 33%]	≤ 0.0001 *
"	"	endocrine/metabolic	28% [19% - 44%]	≤ 0.0001 *
"	"	hematopoietic	35% [24% - 45%]	0.00031 *
"	"	mental disorders	35% [24% - 56%]	0.01
"	"	neurological	35% [22% - 48%]	0.0023
"	"	sense organs	38% [27% - 50%]	0.04
"	"	circulatory system	32% [24% - 43%]	≤ 0.0001 *
"	"	respiratory	33% [26% - 51%]	0.019
"	"	digestive	32% [16% - 50%]	0.00077 *
"	"	genitourinary	41% [30% - 57%]	0.54
"	"	pregnancy complications	47% [37% - 60%]	0.56
"	"	musculoskeletal	47% [33% - 58%]	0.89
"	"	congenital anomalies	25% [18% - 43%]	0.0012
"	"	symptoms	46% [38% - 57%]	0.59
"	"	injuries & poisonings	42% [32% - 55%]	0.22
musculoskeletal	47% [33% - 58%]	infectious diseases	40% [24% - 56%]	0.26
"	"	neoplasms	18% [11% - 33%]	≤ 0.0001 *
"	"	endocrine/metabolic	28% [19% - 44%]	≤ 0.0001 *
"	"	hematopoietic	35% [24% - 45%]	≤ 0.0001 *
"	"	mental disorders	35% [24% - 56%]	0.0049
"	"	neurological	35% [22% - 48%]	0.00083 *
"	"	sense organs	38% [27% - 50%]	0.013
"	"	circulatory system	32% [24% - 43%]	≤ 0.0001 *
"	"	respiratory	33% [26% - 51%]	0.006
"	"	digestive	32% [16% - 50%]	0.0002 *
"	"	genitourinary	41% [30% - 57%]	0.33
"	"	pregnancy complications	47% [37% - 60%]	0.48
"	"	dermatologic	48% [28% - 59%]	0.89
"	"	congenital anomalies	25% [18% - 43%]	0.00063 *
"	"	symptoms	46% [38% - 57%]	0.68

"	"	injuries & poisonings	42% [32% - 55%]	0.13
congenital anomalies	25% [18% - 43%]	infectious diseases	40% [24% - 56%]	0.057
"	"	neoplasms	18% [11% - 33%]	0.0016
"	"	endocrine/metabolic	28% [19% - 44%]	0.82
"	"	hematopoietic	35% [24% - 45%]	0.98
"	"	mental disorders	35% [24% - 56%]	0.29
"	"	neurological	35% [22% - 48%]	0.41
"	"	sense organs	38% [27% - 50%]	0.048
"	"	circulatory system	32% [24% - 43%]	0.6
"	"	respiratory	33% [26% - 51%]	0.12
"	"	digestive	32% [16% - 50%]	0.48
"	"	genitourinary	41% [30% - 57%]	0.003
"	"	pregnancy complications	47% [37% - 60%]	0.043
"	"	dermatologic	48% [28% - 59%]	0.0012
"	"	musculoskeletal	47% [33% - 58%]	0.00063 *
"	"	symptoms	46% [38% - 57%]	0.0018
"	"	injuries & poisonings	42% [32% - 55%]	0.038
symptoms	46% [38% - 57%]	infectious diseases	40% [24% - 56%]	0.16
"	"	neoplasms	18% [11% - 33%]	≤ 0.0001 *
"	"	endocrine/metabolic	28% [19% - 44%]	≤ 0.0001 *
"	"	hematopoietic	35% [24% - 45%]	0.00015 *
"	"	mental disorders	35% [24% - 56%]	0.0052
"	"	neurological	35% [22% - 48%]	0.0016
"	"	sense organs	38% [27% - 50%]	0.015
"	"	circulatory system	32% [24% - 43%]	≤ 0.0001 *
"	"	respiratory	33% [26% - 51%]	0.0049
"	"	digestive	32% [16% - 50%]	0.001
"	"	genitourinary	41% [30% - 57%]	0.21
"	"	pregnancy complications	47% [37% - 60%]	0.34
"	"	dermatologic	48% [28% - 59%]	0.59
"	"	musculoskeletal	47% [33% - 58%]	0.68
"	"	congenital anomalies	25% [18% - 43%]	0.0018
"	"	injuries & poisonings	42% [32% - 55%]	0.1

* *p* < 0.001

APPENDIX F

PheWAS summary under naive post-acute phenotyping: top 100 associations.

Phecode	Description	Odds Ratio (95% CI)	p value	No. cases	No. controls
585.1	Acute renal failure	5.72 (4.91-6.68)	5.07E-109	1,054	183,082
427.5	Arrhythmia (cardiac) NOS	6.85 (5.67-8.26)	1.92E-89	632	178,215
292.4	Altered mental status	12.5 (9.55-16.3)	8.91E-77	258	187,128
418.1	Precordial pain	4.10 (3.53-4.77)	8.73E-75	1,034	183,404
285.1	Acute posthemorrhagic anemia	16.3 (12.0-22.0)	5.43E-73	189	183,114
458.9	Hypotension NOS	10.3 (7.96-13.4)	5.54E-70	274	189,734
401.1	Essential hypertension	1.63 (1.54-1.72)	5.60E-68	15,711	163,382
401.22	Hypertensive chronic kidney disease	6.19 (5.03-7.62)	1.16E-66	738	163,382
285	Other anemias	3.37 (2.93-3.87)	6.74E-65	1,443	183,114
509.1	Respiratory failure	5.07 (4.17-6.16)	7.60E-60	600	188,161
994.2	Sepsis	11.5 (8.50-15.7)	2.06E-55	196	191,158
1013	Asphyxia and hypoxemia	9.14 (6.93-12.1)	2.78E-55	240	190,957
348.8	Encephalopathy, not elsewhere classified	14.6 (10.4-20.4)	5.97E-55	157	184,955
38	Septicemia	12.6 (9.16-17.4)	7.00E-54	176	188,677
772.3	Muscle weakness	5.01 (4.08-6.15)	9.28E-54	536	188,033
272.1	Hyperlipidemia	2.29 (2.06-2.55)	7.06E-53	3,137	171,670
276.6	Fluid overload	8.37 (6.29-11.1)	2.65E-48	245	186,845
512.9	Other dyspnea	2.47 (2.18-2.79)	1.44E-46	1,835	171,917
649	Conditions of the mother complicating pregnancy, childbirth, or the puerperium	5.25 (4.18-6.59)	2.00E-46	368	108,651
276.41	Acidosis	4.87 (3.91-6.05)	6.44E-46	513	186,845
338.1	Acute pain	2.64 (2.30-3.02)	1.07E-43	1,652	177,888
512.7	Shortness of breath	1.95 (1.78-2.15)	1.43E-42	3,451	171,917
250.22	Type 2 diabetes with renal manifestations	3.47 (2.90-4.15)	8.98E-42	1,181	177,510
38.3	Bacteremia	15.1 (10.2-22.4)	1.11E-41	118	188,677
276.5	Hypovolemia	6.93 (5.23-9.17)	1.20E-41	261	186,845
276.12	Hyposmolality and/or hyponatremia	4.89 (3.88-6.17)	3.26E-41	433	186,845
278.11	Morbid obesity	2.10 (1.88-2.35)	2.05E-39	2,335	182,589
260.2	Severe protein-calorie malnutrition	7.91 (5.81-10.8)	2.68E-39	212	183,599
530.11	GERD	1.89 (1.72-2.08)	6.32E-39	3,958	178,019
418	Nonspecific chest pain	2.09 (1.87-2.35)	1.32E-37	2,264	183,404
250.2	Type 2 diabetes	1.85 (1.68-2.04)	1.78E-36	5,316	177,510
401.2	Hypertensive heart and/or renal disease	6.69 (4.98-9.01)	3.48E-36	373	163,382
585.3	Chronic renal failure [CKD]	3.57 (2.91-4.38)	1.94E-34	787	183,082
276.13	Hyperpotassemia	3.56 (2.90-4.38)	1.63E-33	621	186,845
1010.6	Reproductive and maternal health services	2.08 (1.84-2.35)	2.25E-32	1,598	189,380
285.21	Anemia in chronic kidney disease	3.64 (2.94-4.52)	3.35E-32	629	183,114
428.1	Congestive heart failure (CHF) NOS	4.48 (3.49-5.75)	6.83E-32	427	184,871
401.21	Hypertensive heart disease	3.63 (2.92-4.52)	3.78E-31	701	163,382
288.2	Elevated white blood cell count	4.41 (3.43-5.67)	8.23E-31	344	187,148
509.8	Dependence on respirator [Ventilator] or supplemental oxygen	7.42 (5.24-10.5)	1.37E-29	160	188,161
665	Obstetrical/birth trauma	7.96 (5.54-11.4)	2.50E-29	152	191,347
411.2	Myocardial infarction	3.44 (2.75-4.31)	5.97E-27	607	182,562
327.32	Obstructive sleep apnea	1.75 (1.58-1.94)	7.65E-27	3,401	179,431
508	Pulmonary collapse; interstitial and compensatory emphysema	6.58 (4.66-9.29)	8.33E-27	176	188,161
426	Cardiac conduction disorders	9.15 (5.95-14.1)	6.94E-24	106	178,215
591	Urinary tract infection	2.29 (1.95-2.69)	1.75E-23	1,145	184,472
798	Malaise and fatigue	1.63 (1.48-1.80)	2.00E-23	3,445	178,495
276.14	Hypopotassemia	3.36 (2.65-4.27)	2.91E-23	466	186,845

260.3	Adult failure to thrive	8.99 (5.81-13.9)	5.90E-23	101	183,599
789	Nausea and vomiting	1.83 (1.62-2.07)	1.44E-22	2,139	184,375
994.21	Septic shock	18.0 (10.1-32.4)	2.70E-22	53	191,158
797	Shock	16.2 (9.16-28.5)	7.79E-22	55	191,476
297.1	Suicidal ideation	5.50 (3.89-7.80)	7.90E-22	192	170,204
655	Known or suspected fetal abnormality affecting management of mother	3.41 (2.65-4.38)	1.15E-21	310	108,936
272.11	Hypercholesterolemia	1.82 (1.61-2.06)	1.24E-21	2,281	171,670
244.4	Hypothyroidism NOS	1.69 (1.51-1.88)	4.88E-21	2,802	182,090
250.24	Type 2 diabetes with neurological manifestations	2.29 (1.93-2.73)	5.53E-21	1,344	177,510
1010	Other tests	3.50 (2.69-4.56)	1.08E-20	308	189,547
427.7	Tachycardia NOS	2.55 (2.10-3.11)	1.41E-20	663	178,215
646	Other complications of pregnancy NEC	5.10 (3.60-7.22)	3.68E-20	160	109,219
278.1	Obesity	1.90 (1.66-2.18)	3.98E-20	1,549	182,589
644	Anemia during pregnancy	10.0 (6.09-16.6)	1.62E-19	73	109,431
274.1	Gout	3.25 (2.51-4.21)	3.73E-19	351	189,924
654.1	Abnormality of organs and soft tissues of pelvis complicating pregnancy, childbirth, or the puerperium	4.04 (2.97-5.50)	5.69E-19	209	108,966
136	Other infectious and parasitic diseases	6.74 (4.41-10.3)	1.17E-18	104	191,038
428.4	Heart failure with preserved EF [Diastolic heart failure]	2.39 (1.96-2.91)	7.53E-18	1,167	184,871
663	Umbilical cord complications during labor and delivery	21.8 (10.8-44.1)	1.02E-17	44	191,521
638	Other high-risk pregnancy	2.38 (1.95-2.90)	1.06E-17	584	190,705
290.2	Delirium due to conditions classified elsewhere	20.4 (10.2-40.6)	1.07E-17	37	187,128
411.4	Coronary atherosclerosis	1.60 (1.44-1.79)	2.26E-17	4,362	182,562
585.32	End stage renal disease	2.69 (2.14-3.38)	2.37E-17	722	183,082
588.1	Renal osteodystrophy	3.12 (2.39-4.06)	3.32E-17	456	183,082
427.21	Atrial fibrillation	1.71 (1.51-1.94)	5.90E-17	3,184	178,215
284.1	Pancytopenia	3.94 (2.85-5.45)	1.01E-16	256	183,114
785	Abdominal pain	1.42 (1.31-1.54)	1.27E-16	5,027	180,428
452.2	Deep vein thrombosis [DVT]	2.57 (2.05-3.21)	1.63E-16	569	185,983
480	Pneumonia	2.83 (2.21-3.63)	1.81E-16	387	188,133
653	Problems associated with amniotic cavity and membranes	13.4 (7.22-24.8)	1.86E-16	48	108,966
585.34	Chronic Kidney Disease, Stage IV	2.91 (2.26-3.76)	2.25E-16	537	183,082
590	Pyelonephritis	5.81 (3.81-8.86)	2.95E-16	113	184,472
851	Complications of transplants and reattached limbs	2.38 (1.94-2.94)	3.18E-16	788	187,602
287.3	Thrombocytopenia	2.99 (2.30-3.90)	5.75E-16	382	188,946
507	Pleurisy; pleural effusion	3.24 (2.43-4.31)	7.34E-16	351	188,161
280.1	Iron deficiency anemias, unspecified or not due to blood loss	1.91 (1.63-2.24)	1.36E-15	1,330	183,114
578.9	Hemorrhage of gastrointestinal tract	5.45 (3.59-8.29)	2.10E-15	123	188,746
260	Protein-calorie malnutrition	3.39 (2.50-4.58)	2.90E-15	302	183,599
41	Bacterial infection NOS	3.15 (2.36-4.21)	5.79E-15	288	188,677
359.2	Myopathy	8.12 (4.79-13.8)	7.43E-15	71	188,619
285.2	Anemia of chronic disease	6.90 (4.24-11.2)	8.32E-15	89	183,114
250.42	Other abnormal glucose	2.16 (1.78-2.63)	8.87E-15	805	177,510
272.13	Mixed hyperlipidemia	1.44 (1.31-1.58)	8.95E-15	5,236	171,670
38.2	Gram positive septicemia	16.5 (8.08-33.7)	1.43E-14	36	188,677
41.4	E. coli	13.3 (6.86-25.9)	2.25E-14	41	188,677
41.9	Infection with drug-resistant microorganisms	15.1 (7.53-30.3)	2.26E-14	37	188,677
1090	Acquired absence of organs	2.55 (2.00-3.24)	2.80E-14	569	189,321
642	Hypertension complicating pregnancy, childbirth, and the puerperium	5.51 (3.53-8.61)	6.49E-14	93	109,399
567	Peritonitis and retroperitoneal infections	4.84 (3.20-7.32)	8.48E-14	135	187,178
276.11	Hyperosmolality and/or hypernatremia	34.2 (13.4-87.6)	1.65E-13	22	186,845
532	Dysphagia	1.84 (1.56-2.16)	4.40E-13	1,445	178,019

APPENDIX G

Subgroup analyses of temporal-informed PheWAS

Temporal-informed PheWAS by race/ethnicity subgroups.

White, Non-Hispanic patients (N=125,938)						
Phecode[a]	**Description**	**Odds Ratio**	**95% CI**	**p value**	**No. cases**	**No. controls**
512.9	Other dyspnea	3.22	(2.59-4.00)	7.73E-26	627	59,636
512.7	Shortness of breath	2.81	(2.29-3.44)	1.66E-23	760	59,636
359.2	Myopathy	29.9	(12.2-73.3)	1.26E-13	25	113,531
427.9	Palpitations	2.35	(1.85-2.97)	1.26E-12	468	84,577
569.2	Gastrointestinal complications of surgery	5.69	(3.48-9.30)	3.93E-12	85	107,583
418.1	Precordial pain	3.78	(2.59-5.52)	5.31E-12	194	87,353
278.11	Morbid obesity	2.32	(1.82-2.96)	1.08E-11	437	100,068
136	Other infectious and parasitic diseases	11.4	(5.65-23.0)	1.14E-11	38	119,882
509.1	Respiratory failure	6.32	(3.66-10.9)	3.92E-11	86	100,780
649	Conditions of the mother complicating pregnancy, childbirth, or the puerperium	3.77	(2.53-5.62)	6.23E-11	124	63,357
427.21	Atrial fibrillation	2.67	(1.95-3.64)	6.86E-10	382	84,577
585.1	Acute renal failure	3.25	(2.18-4.86)	8.87E-09	218	100,977
327.32	Obstructive sleep apnea	2.05	(1.60-2.62)	1.14E-08	526	94,165
418	Nonspecific chest pain	1.97	(1.55-2.50)	2.52E-08	530	87,353
782.3	Edema	2.27	(1.67-3.08)	1.78E-07	331	108,722
1010	Other tests	3.26	(2.09-5.11)	2.23E-07	111	110,956
250.2	Type 2 diabetes	2.13	(1.60-2.84)	2.41E-07	406	93,983
599.2	Retention of urine	3.39	(2.13-5.41)	2.95E-07	148	94,495
1013	Asphyxia and hypoxemia	6.28	(3.11-12.7)	3.09E-07	43	114,386
350.1	Abnormal involuntary movements	2.55	(1.76-3.69)	6.66E-07	205	111,360
249	Secondary diabetes mellitus	8.71	(3.62-21.0)	1.39E-06	44	93,983
452.2	Deep vein thrombosis [DVT]	3.45	(2.08-5.74)	1.76E-06	105	104,177
452	Other venous embolism and thrombosis	4.49	(2.41-8.37)	2.22E-06	68	104,177
420.1	Myocarditis	11.4	(4.05-31.9)	3.84E-06	17	115,711
292	Neurological disorders	2.47	(1.68-3.62)	3.90E-06	188	104,296
514	Abnormal findings examination of lungs	2.42	(1.66-3.53)	4.80E-06	284	105,200
646	Other complications of pregnancy NEC	4.09	(2.22-7.51)	5.71E-06	50	65,813
1010.6	Reproductive and maternal health services	1.71	(1.35-2.17)	8.43E-06	408	114,589
284.1	Pancytopenia	3.91	(2.11-7.22)	1.38E-05	73	94,032
727.1	Synovitis and tenosynovitis	2.31	(1.58-3.37)	1.47E-05	212	92,564
386.9	Dizziness and giddiness	1.74	(1.35-2.25)	1.69E-05	489	103,688
781	Symptoms involving nervous and musculoskeletal systems	3.01	(1.82-4.98)	1.83E-05	117	117,833
433.1	Occlusion and stenosis of precerebral arteries	3.66	(1.99-6.73)	3.16E-05	83	111,257
401.1	Essential hypertension	1.45	(1.22-1.73)	3.38E-05	1,268	74,057
285.2	Anemia of chronic disease	11.0	(3.53-34.2)	3.54E-05	17	94,032
649.1	Diabetes or abnormal glucose tolerance complicating pregnancy	4.61	(2.21-9.61)	4.58E-05	36	63,357
Black, Non-Hispanic patients (N=19,936)						
798.1	Chronic fatigue syndrome	7.63	(3.39-17.2)	8.98E-07	29	12,789
569.2	Gastrointestinal complications of surgery	8.13	(3.50-18.9)	1.07E-06	24	16,681
278.11	Morbid obesity	2.95	(1.87-4.65)	3.43E-06	106	13,894
638	Other high-risk pregnancy	4.38	(2.19-8.78)	3.07E-05	37	18,235
250.42	Other abnormal glucose	4.76	(2.23-10.1)	5.51E-05	42	13,745
280.2	Iron deficiency anemia secondary to blood loss	5.08	(2.20-11.7)	1.39E-04	28	12,275

[a] A list of ICD-10-CM codes included in each phecode is available at: https://phewascatalog.org/phecodes_icd10cm [22]

Temporal-informed PheWAS by sex.

Temporal-informed PheWAS among both male and female patients.

		Females					Males				
Phenotype	**Description**	**Odds Ratio**	**95% CI**	**p value**	**No. cases**	**No. controls**	**Odds Ratio**	**95% CI**	**p value**	**No. cases**	**No. controls**
512.9	Other dyspnea	2.75	(2.13-3.55)	9.38×10^{-15}	456	52,865	3.46	(2.61-4.58)	6.56×10^{-18}	355	41,071
512.7	Shortness of breath	2.24	(1.79-2.82)	3.35×10^{-12}	597	52,865	2.94	(2.22-3.88)	3.81×10^{-14}	391	41,071
278.11	Morbid obesity	2.29	(1.82-2.87)	1.06×10^{-12}	464	85,699	2.50	(1.7-3.69)	3.76×10^{-06}	160	69,162
359.2	Myopathy	25.2	(7.23-88.2)	4.18×10^{-07}	15	100,096	18.0	(6.24-51.9)	8.97×10^{-08}	18	74,767
418.1	Precordial pain	2.99	(2-4.47)	8.81×10^{-08}	174	79,908	3.66	(2.2-6.07)	5.16×10^{-07}	104	58,629
418	Nonspecific chest pain	1.93	(1.51-2.47)	1.46×10^{-07}	444	79,908	2.13	(1.56-2.9)	1.84×10^{-06}	302	58,629
585.1	Acute renal failure	3.61	(2.27-5.76)	6.15×10^{-08}	151	92,868	2.76	(1.72-4.42)	2.47×10^{-05}	158	64,607

Temporal-informed PheWAS among female patients only.

		Females					Males				
Phenotype	**Description**	**Odds Ratio**	**95% CI**	**p value**	**No. cases**	**No. controls**	**Odds Ratio**	**95% CI**	**p value**	**No. cases**	**No. controls**
569.2	Gastrointestinal complications	6.54	(4.24-10.1)	2.51×10^{-17}	94	95,244	5.78	(1.96-17)	0.0014	22	71,581
136	Other infectious and parasitic diseases	16.0	(7.57-33.7)	3.55×10^{-13}	35	103,699	3.33	(1.19-9.26)	0.0214	19	78,267
427.9	Palpitations	2.35	(1.86-2.97)	1.07×10^{-12}	441	80,047	1.67	(1.12-2.47)	0.011	187	57,039
649	Conditions of the mother complicating pregnancy, childbirth, or the puerperium	2.94	(2.15-4.03)	1.71×10^{-11}	169	95,518	-	-	-	-	-
646	Other complications of pregnancy NEC	4.27	(2.66-6.86)	2.06×10^{-09}	69	99,542	-	-	-	-	-
1010.6	Reproductive and maternal health services	1.75	(1.44-2.12)	9.99×10^{-09}	591	92,952	-	-	-	-	-
638	Other high-risk pregnancy	2.19	(1.67-2.86)	1.34×10^{-08}	312	98,919	-	-	-	-	-
1010	Other tests	3.19	(2.09-4.86)	7.43×10^{-08}	116	92,964	2.7	(1.22-5.95)	0.014	39	76,383
420.1	Myocarditis	16.1	(5.52-47)	3.67×10^{-07}	15	102,140	-	-	-	5	74,863
401.1	Essential hypertension	1.63	(1.34-1.98)	1.01×10^{-06}	865	74,906	1.22	(0.98-1.52)	0.0725	833	48,001
644	Anemia during pregnancy	4.72	(2.5-8.92)	1.81×10^{-06}	38	101,761	-	-	-	-	-
452.2	Deep vein thrombosis [DVT]	4.37	(2.38-8.03)	1.96×10^{-06}	66	93,371	2.37	(1.27-4.44)	0.007	72	69,340
285	Other anemias	2.34	(1.64-3.32)	2.21×10^{-06}	259	82,899	1.69	(1.11-2.58)	0.0149	214	63,606
649.1	Diabetes or abnormal glucose tolerance complicating pregnancy	3.52	(2.08-5.98)	3.11×10^{-06}	57	95,518	-	-	-	-	-
292	Neurological disorders	2.70	(1.76-4.14)	5.11×10^{-06}	135	93,719	2.01	(1.2-3.37)	0.00826	107	68,515
671	Venous/cerebrovascular complications embolism in pregnancy and the puerperium	9.77	(3.61-26.4)	7.09×10^{-06}	17	103,586	-	-	-	-	-

782.3	Edema	2.25	(1.57-3.22)	9.51×10^{-06}	231	95,466	1.91	(1.27-2.88)	0.00182	193	72,718
348.8	Encephalopathy, not elsewhere classified	12.1	(3.95-37.1)	1.27×10^{-05}	17	92,273	3.40	(0.91-12.7)	0.0694	15	68,246
284.1	Pancytopenia	5.01	(2.42-10.4)	1.45×10^{-05}	46	82,899	2.02	(0.84-4.84)	0.114	48	63,606

Temporal-informed PheWAS among male patients only.											
		Females					**Males**				
Phenotype	**Description**	**Odds Ratio**	**95% CI**	**p value**	**No. cases**	**No. controls**	**Odds Ratio**	**95% CI**	**p value**	**No. cases**	**No. controls**
509.1	Respiratory failure	3.23	(1.45-7.17)	0.00401	50	93,041	12.4	(6.53-23.5)	1.27×10^{-14}	51	64,751
427.21	Atrial fibrillation	1.96	(1.23-3.13)	0.00462	170	80,047	3.27	(2.28-4.68)	9.13×10^{-11}	273	57,039
840	Sprains and strains	1.10	(0.80-1.51)	0.57	348	91,895	2.28	(1.64-3.18)	1.08×10^{-06}	249	69,749
514	Abnormal findings examination of lungs	1.76	(1.09-2.82)	0.02	183	95,645	3.12	(1.95-5)	2.04×10^{-06}	167	67,924
278.1	Obesity	1.41	(1.07-1.85)	0.015	401	85,699	2.41	(1.65-3.51)	5.73×10^{-06}	165	69,162
350.1	Abnormal involuntary movements	2.22	(1.44-3.43)	0.0003	149	96,901	2.99	(1.85-4.84)	8.5×10^{-06}	107	73,586
1013	Asphyxia and hypoxemia	4.40	(1.73-11.2)	0.002	27	101,438	7.13	(2.89-17.6)	1.98×10^{-05}	25	74,001
295.1	Schizophrenia	-	-	-	9	66,054	6.71	(2.67-16.9)	5.13×10^{-05}	26	58,256

Temporal-informed PheWAS by onset of diagnosis.

Phenotype	Description	Early-presenting post-acute phenotypes [a]					Late-presenting post-acute phenotypes [b]				
		Odds Ratio	95% CI	p value	No. cases	No. controls	Odds Ratio	95% CI	p value	No. cases	No. controls
512.9	Other dyspnea	3.89	(2.80-5.41)	7.14E-16	227	96,877	2.79	(2.22-3.51)	1.13E-18	584	90,181
569.2	Gastrointestinal complications	9.81	(5.21-18.5)	1.62E-12	44	167,795	5.11	(3.06-8.53)	4.31E-10	72	160,967
509.1	Respiratory failure	12.8	(6.24-26.2)	3.25E-12	43	158,293	4.42	(2.22-8.82)	2.38E-05	58	152,379
427.9	Palpitations	3.02	(2.11-4.32)	1.68E-09	172	138,826	1.86	(1.46-2.37)	6.56E-07	456	132,806
512.7	Shortness of breath	2.52	(1.86-3.42)	2.83E-09	292	96,877	2.52	(2.03-3.11)	2.05E-17	696	90,181
644	Anemia during pregnancy	19.0	(6.26-57.4)	1.97E-07	16	101,815	-	-	-	-	-
514	Abnormal findings examination of lungs	3.47	(2.06-5.85)	2.84E-06	113	164,532	-	-	-	-	-
285	Other anemias	2.83	(1.83-4.39)	3.41E-06	157	147,698	-	-	-	-	-
418.1	Precordial pain	4.42	(2.34-8.37)	4.74E-06	56	140,112	2.94	(2.05-4.21)	4.7E-09	222	134,057
136	Other infectious and parasitic diseases	11.9	(3.81-37.1)	2.03E-05	15	182,227	8.46	(4.26-16.8)	1.03E-09	39	175,181
585.1	Acute renal failure	3.24	(1.84-5.69)	4.69E-05	107	158,429	3.26	(2.19-4.85)	5.22E-09	202	152,145
359.2	Myopathy	12.7	(3.74-43.5)	4.78E-05	16	175,579	28.1	(9.94-79.6)	3.23E-10	17	168,524
452	Other venous embolism and thrombosis	5.49	(2.36-12.7)	7.37E-05	32	164,545	-	-	-	-	-
401.1	Essential hypertension	1.62	(1.27-2.06)	9.52E-05	617	124,813	-	-	-	-	-
278.11	Morbid obesity	-	-	-	-	-	2.49	(1.98-3.14)	9.64E-15	443	149,748
649	Conditions of the mother complicating pregnancy, childbirth, or the puerperium	-	-	-	-	-	3.66	(2.54-5.27)	3.26E-12	145	92,130
427.21	Atrial fibrillation	-	-	-	-	-	3.23	(2.30-4.56)	1.88E-11	263	132,806
418	Nonspecific chest pain	-	-	-	-	-	1.99	(1.59-2.49)	1.41E-09	566	134,057
292	Neurological disorders	-	-	-	-	-	2.97	(2.09-4.22)	1.51E-09	192	156,592
646	Other complications of pregnancy NEC	-	-	-	-	-	5.06	(2.86-8.93)	2.3E-08	55	95,983
1010	Other tests	-	-	-	-	-	3.23	(2.11-4.94)	7.01E-08	121	163,187
638	Other high-risk pregnancy	-	-	-	-	-	2.23	(1.66-2.98)	8.46E-08	269	172,113
395.6	Heart valve replaced	-	-	-	-	-	11.0	(4.43-27.5)	2.53E-07	23	160,821
1010.6	Reproductive and maternal health services	-	-	-	-	-	1.74	(1.41-2.15)	2.55E-07	496	166,455
350.1	Abnormal involuntary movements	-	-	-	-	-	2.61	(1.81-3.77)	3.19E-07	185	164,415
782.3	Edema	-	-	-	-	-	2.22	(1.62-3.03)	5.54E-07	301	162,355
781	Symptoms involving nervous and musculoskeletal systems	-	-	-	-	-	3.44	(2.11-5.60)	7.36E-07	105	173,467
433.1	Occlusion and stenosis of precerebral arteries	-	-	-	-	-	4.72	(2.47-9.00)	2.6E-06	64	165,557
260	Protein-calorie malnutrition	-	-	-	-	-	4.23	(2.29-7.82)	4.06E-06	69	151,935
278.1	Obesity	-	-	-	-	-	1.80	(1.40-2.32)	4.28E-06	437	149,748
348.8	Encephalopathy, not elsewhere classified	-	-	-	-	-	11.6	(4.06-33.0)	4.6E-06	19	154,973
649.1	Diabetes or abnormal glucose tolerance complicating pregnancy	-	-	-	-	-	4.91	(2.48-9.72)	5.01E-06	39	92,130

79	Viral infection	-	-	-	-	-	3.97	(2.18-7.23)	6.42E-06	74	144,377
244.4	Hypothyroidism NOS	-	-	-	-	-	2.10	(1.51-2.92)	1.04E-05	261	154,957
327.32	Obstructive sleep apnea	-	-	-	-	-	1.77	(1.37-2.29)	1.39E-05	479	145,658
276.13	Hyperpotassemia	-	-	-	-	-	3.18	(1.87-5.41)	1.96E-05	108	144,713
199	Neoplasm of uncertain behavior	-	-	-	-	-	3.23	(1.87-5.60)	2.83E-05	84	166,548
428.4	Heart failure with preserved EF	-	-	-	-	-	2.78	(1.69-4.56)	5.17E-05	142	159,733
285.22	Anemia in neoplastic disease	-	-	-	-	-	2.88	(1.73-4.82)	5.18E-05	117	141,744
701.2	Scar conditions and fibrosis of skin	-	-	-	-	-	3.04	(1.77-5.23)	5.44E-05	102	166,637

[a] First post-acute diagnosis made within 60 days of recovery following COVID-19 testing

[b] First post-acute diagnosis made after 60 days of recovery following COVID-19 testing

APPENDIX H

Sensitivity analyses results

Sensitivity analysis: adults with a least one follow-up visit in post-acute period.

Phecode	Description	Odds Ratio	95% CI	p value	No. cases	No. controls
512.9	Other dyspnea	3.05	(2.53-3.68)	3.18E-31	811	50,602
512.7	Shortness of breath	2.52	(2.11-3.00)	6.45E-25	988	50,602
569.2	Gastrointestinal complications	5.86	(3.94-8.73)	3.00E-18	116	97,430
278.11	Morbid obesity	2.24	(1.84-2.72)	6.34E-16	623	88,502
509.1	Respiratory failure	6.99	(4.28-11.4)	6.96E-15	101	93,258
649	Conditions of the mother complicating pregnancy, childbirth, or the puerperium	3.65	(2.61-5.10)	3.49E-14	167	60,558
359.2	Myopathy	19.1	(8.69-41.9)	2.03E-13	33	103,726
136	Other infectious and parasitic diseases	8.63	(4.84-15.4)	2.73E-13	54	110,018
418.1	Precordial pain	3.03	(2.22-4.13)	2.7E-12	278	78,592
427.9	Palpitations	2.03	(1.66-2.48)	5.26E-12	627	76,834
418	Nonspecific chest pain	1.95	(1.61-2.36)	8.73E-12	745	78,592
646	Other complications of pregnancy NEC	5.35	(3.23-8.87)	8.21E-11	69	63,503
427.21	Atrial fibrillation	2.53	(1.91-3.36)	1.29E-10	443	76,834
585.1	Acute renal failure	2.93	(2.11-4.08)	1.59E-10	309	91,680
1010.6	Reproductive and maternal health services	1.79	(1.47-2.19)	6.06E-09	549	104,034
1010	Other tests	2.97	(2.05-4.29)	8.00E-09	155	99,632
644	Anemia during pregnancy	6.69	(3.37-13.3)	5.62E-08	38	65,119
350.1	Abnormal involuntary movements	2.40	(1.74-3.30)	7.54E-08	256	101,168
671	Venous/cerebrovascular complications embolism in pregnancy and the puerperium	19.2	(6.51-56.6)	8.6E-08	17	66,122
638	Other high-risk pregnancy	2.11	(1.61-2.78)	9.57E-08	298	107,781
452.2	Deep vein thrombosis [DVT]	3.11	(2.02-4.79)	2.61E-07	138	93,807
649.1	Diabetes or abnormal glucose tolerance complicating pregnancy	4.40	(2.50-7.75)	2.68E-07	57	60,558
292	Neurological disorders	2.33	(1.68-3.22)	4.11E-07	242	96,775
1013	Asphyxia and hypoxemia	5.24	(2.76-9.95)	4.28E-07	52	105,514
782.3	Edema	1.99	(1.52-2.60)	4.72E-07	424	98,285
781	Symptoms involving nervous and musculoskeletal systems	2.89	(1.90-4.39)	7.16E-07	151	108,474
599.2	Retention of urine	2.72	(1.82-4.08)	1.21E-06	184	82,558
285	Other anemias	1.92	(1.47-2.51)	1.87E-06	473	83,038
401.1	Essential hypertension	1.42	(1.23-1.64)	2.71E-06	1697	66,158
587	Kidney replaced by transpant	26.5	(6.61-106.)	3.71E-06	22	91,680
514	Abnormal findings examination of lungs	2.18	(1.57-3.04)	3.75E-06	350	96,271
420.1	Myocarditis	9.03	(3.47-23.5)	6.64E-06	20	105,934
250.2	Type 2 diabetes	1.73	(1.36-2.20)	9.29E-06	572	83,868
278.1	Obesity	1.64	(1.31-2.04)	1.07E-05	565	88,502
327.32	Obstructive sleep apnea	1.62	(1.30-2.01)	1.54E-05	669	84,158
348.8	Encephalopathy, not elsewhere classified	5.96	(2.64-13.5)	1.79E-05	32	94,836
502	Postinflammatory pulmonary fibrosis	5.34	(2.44-11.7)	2.79E-05	40	93,258
653	Problems associated with amniotic cavity and membranes	7.02	(2.75-17.9)	4.49E-05	19	61,980
655	Known or suspected fetal abnormality affecting management of mother	2.13	(1.48-3.06)	4.55E-05	158	59,807

Sensitivity analysis: phenotype case definition of one post-acute phecode.

Phecode	Description	Odds Ratio	95% CI	p value	No. cases	No. controls
512.9	Other dyspnea	2.13	(1.89-2.41)	8.88E-35	2202	93,936
136	Other infectious and parasitic diseases	6.51	(4.74-8.92)	3.24E-31	185	181,966
418.1	Precordial pain	2.62	(2.19-3.14)	6.66E-26	978	138,537
509.1	Respiratory failure	4.85	(3.61-6.50)	8.38E-26	304	157,792
278.11	Morbid obesity	1.81	(1.59-2.05)	3.45E-20	1678	154,861
646	Other complications of pregnancy NEC	3.61	(2.73-4.78)	3.16E-19	240	99,542
512.7	Shortness of breath	1.60	(1.44-1.77)	5.86E-19	3182	93,936
649	Conditions of the mother complicating pregnancy, childbirth, or the puerperium	2.97	(2.32-3.79)	3.25E-18	323	95,518
647	Infectious and parasitic complications affecting pregnancy	12.33	(6.84-22.2)	7.03E-17	65	102,065
636	Early or threatened labor; hemorrhage in early pregnancy	7.35	(4.59-11.8)	8.94E-17	79	95,926
1010	Other tests	1.77	(1.55-2.03)	1.92E-16	1365	169,347
569.2	Gastrointestinal complications	3.83	(2.77-5.29)	3.59E-16	221	166,825
285	Other anemias	1.88	(1.61-2.20)	7.55E-16	1515	146,505
585.1	Acute renal failure	2.47	(1.98-3.08)	1.16E-15	792	157,475
427.21	Atrial fibrillation	2.37	(1.90-2.96)	1.52E-14	793	137,086
348.8	Encephalopathy, not elsewhere classified	5.32	(3.40-8.34)	2.85E-13	113	160,519
359.2	Myopathy	5.81	(3.62-9.32)	3.38E-13	91	174,863
782.3	Edema	1.76	(1.51-2.04)	3.47E-13	1396	168,184
644	Anemia during pregnancy	4.62	(3.00-7.10)	3.19E-12	111	101,761
327.32	Obstructive sleep apnea	1.63	(1.42-1.87)	3.47E-12	1737	150,608
798.1	Chronic fatigue syndrome	1.81	(1.53-2.15)	4.63E-12	946	131,770
418	Nonspecific chest pain	1.49	(1.33-1.67)	4.65E-12	2233	138,537
427.5	Arrhythmia (cardiac) NOS	2.13	(1.72-2.64)	5.6E-12	870	137,086
427.9	Palpitations	1.65	(1.43-1.92)	1.45E-11	1306	137,086
278.1	Obesity	1.56	(1.37-1.77)	1.49E-11	1802	154,861
452.2	Deep vein thrombosis [DVT]	2.7	(2.01-3.62)	4.58E-11	327	162,711
292	Neurological disorders	1.97	(1.61-2.41)	4.75E-11	668	162,234
671	Venous/cerebrovascular complications embolism in pregnancy and the puerperium	8.74	(4.54-16.8)	9.23E-11	45	103,586
772.3	Muscle weakness	1.79	(1.49-2.14)	2.86E-10	1139	155,495
661	Fetal distress and abnormal forces of labor	9.70	(4.70-20.0)	7.75E-10	34	181,989
401.1	Essential hypertension	1.36	(1.23-1.49)	8.65E-10	4049	122,907
286.7	Other and unspecified coagulation defects	2.76	(1.98-3.86)	2.63E-09	283	168,886
514	Abnormal findings examination of lungs	1.65	(1.40-1.94)	2.84E-09	1442	163,569
1013	Asphyxia and hypoxemia	2.90	(2.04-4.12)	3.2E-09	215	175,439
655	Known or suspected fetal abnormality affecting management of mother	1.94	(1.55-2.43)	8.5E-09	433	94,447
509.2	Respiratory insufficiency	8.21	(3.97-17.0)	1.27E-08	40	157,792
260	Protein-calorie malnutrition	2.26	(1.70-3.02)	2.48E-08	438	157,175
642	Hypertension complicating pregnancy, childbirth, and the puerperium	3.68	(2.31-5.86)	4.35E-08	90	102,574
420.1	Myocarditis	7.72	(3.69-16.1)	5.43E-08	36	177,003
288.2	Elevated white blood cell count	1.96	(1.54-2.50)	6.43E-08	503	158,244
638	Other high-risk pregnancy	1.85	(1.48-2.31)	8.07E-08	481	178,757
599.3	Dysuria	1.35	(1.21-1.50)	8.61E-08	2383	149,134
727.1	Synovitis and tenosynovitis	1.79	(1.44-2.21)	8.76E-08	688	148,134
276.6	Fluid overload	3.30	(2.13-5.12)	9.52E-08	191	149,733
595	Hydronephrosis	2.45	(1.75-3.44)	1.97E-07	290	176,011
250.2	Type 2 diabetes	1.55	(1.31-1.83)	2.25E-07	1418	148,033

427.7	Tachycardia NOS	1.62	(1.35-1.95)	2.71E-07	857	137,086
38.3	Bacteremia	4.25	(2.44-7.41)	3.32E-07	73	166,009
704.1	Alopecia	1.74	(1.41-2.15)	3.33E-07	563	174,033
649.1	Diabetes or abnormal glucose tolerance complicating pregnancy	3.23	(2.04-5.12)	5.51E-07	96	95,518
260.2	severe protein-calorie malnutrition	3.03	(1.95-4.68)	6.77E-07	165	157,175
790.6	Other abnormal blood chemistry	1.36	(1.21-1.54)	7.51E-07	2209	168,380
535.2	Atrophic gastritis	1.76	(1.40-2.20)	8.30E-07	684	170,481
38	Septicemia	4.68	(2.53-8.68)	9.48E-07	63	166,009
642.1	Preeclampsia and eclampsia	5.74	(2.84-11.6)	1.1E-06	46	102,574
1009	Injury, NOS	1.32	(1.18-1.47)	1.16E-06	2374	148,125
411.2	Myocardial infarction	1.99	(1.51-2.63)	1.24E-06	542	159,985
550.2	Diaphragmatic hernia	1.69	(1.37-2.10)	1.27E-06	787	170,725
306	Other mental disorder	2.43	(1.70-3.49)	1.36E-06	193	124,310
745	Pain in joint	1.24	(1.14-1.35)	1.46E-06	4298	133,022
455	Hemorrhoids	1.41	(1.23-1.62)	1.55E-06	1724	162,711
593	Hematuria	1.57	(1.31-1.89)	1.7E-06	861	152,649
617	Disorders secondary to childbirth, surgery, trauma	1.50	(1.27-1.77)	1.99E-06	934	181,721
994.2	Sepsis	4.06	(2.28-7.25)	2.03E-06	73	178,584
411.4	Coronary atherosclerosis	1.53	(1.29-1.83)	2.1E-06	1374	159,985
647.1	Infections of genitourinary tract during pregnancy	4.11	(2.29-7.36)	2.11E-06	58	102,065
599.2	Retention of urine	1.85	(1.43-2.38)	2.23E-06	527	149,134
272.1	Hyperlipidemia	1.38	(1.21-1.58)	2.96E-06	2322	140,288
567	Peritonitis and retroperitoneal infections	3.53	(2.08-5.99)	2.98E-06	95	166,825
619.4	Noninflammatory disorders of vagina	1.50	(1.26-1.78)	3.04E-06	917	93,150
781	Symptoms involving nervous and musculoskeletal systems	1.92	(1.46-2.53)	3.76E-06	430	180,070
480.2	Viral pneumonia	6.74	(3.00-15.2)	3.78E-06	31	153,134
350.1	Abnormal involuntary movements	1.6	(1.31-1.95)	3.83E-06	807	170,487
502	Postinflammatory pulmonary fibrosis	2.48	(1.68-3.65)	4.54E-06	201	157,792
244.4	Hypothyroidism NOS	1.49	(1.26-1.77)	4.58E-06	1163	160,570
292.4	Altered mental status	2.13	(1.54-2.94)	4.60E-06	315	162,234
656	Other perinatal conditions of fetus or newborn	2.4	(1.65-3.51)	5.26E-06	149	182,164
509.8	Dependence on respirator [Ventilator] or supplemental oxygen	3.24	(1.94-5.41)	7.34E-06	123	157,792
41.9	Infection with drug-resistant microorganisms	6.52	(2.87-14.8)	7.48E-06	30	166,009
401.22	Hypertensive chronic kidney disease	1.98	(1.45-2.69)	1.34E-05	582	122,907
773	Pain in limb	1.23	(1.12-1.36)	1.45E-05	3532	147,087
261.4	Vitamin D deficiency	1.35	(1.18-1.54)	1.77E-05	1782	157,175
395.2	Nonrheumatic aortic valve disorders	1.88	(1.41-2.50)	1.8E-05	412	166,630
1010.6	Reproductive and maternal health services	1.43	(1.21-1.68)	2.04E-05	865	172,787
304	Adjustment reaction	1.47	(1.23-1.76)	2.27E-05	899	124,310
357	Inflammatory and toxic neuropathy	1.51	(1.25-1.83)	2.32E-05	997	174,863
532	Dysphagia	1.46	(1.23-1.75)	2.38E-05	1155	137,590
427.22	Atrial flutter	2.23	(1.54-3.23)	2.40E-05	320	137,086
300.1	Anxiety disorder	1.22	(1.11-1.34)	2.41E-05	3961	124,310
250.42	Other abnormal glucose	1.40	(1.20-1.64)	2.83E-05	1548	148,033
386.9	Dizziness and giddiness (Light-headedness and vertigo)	1.30	(1.15-1.48)	2.84E-05	2066	161,499
250.22	Type 2 diabetes with renal manifestations	2.34	(1.57-3.48)	3.08E-05	496	148,033
587	Kidney replaced by transpant	5.42	(2.44-12.1)	3.38E-05	56	157,475
504	Other alveolar and parietoalveolar pneumonopathy	2.72	(1.69-4.38)	3.56E-05	120	157,792

Sensitivity analysis: adults with ≥6 months of EHR records prior to testing.

Phecode	Description	Odds Ratio	95% CI	p value	No. cases	No. controls
512.9	Other dyspnea	2.80	(2.27-3.46)	1.16E-21	680	53,761
512.7	Shortness of breath	2.45	(2.02-2.98)	1.28E-19	825	53,761
569.2	Gastrointestinal complications	7.06	(4.47-11.2)	5.48E-17	87	110,991
278.11	Morbid obesity	2.37	(1.91-2.93)	4.59E-15	506	101,668
136	Other infectious and parasitic diseases	10.0	(5.48-18.4)	8.24E-14	50	123,943
418.1	Precordial pain	3.23	(2.34-4.47)	1.39E-12	254	86,891
649	Conditions of the mother complicating pregnancy, childbirth, or the puerperium	3.56	(2.42-5.22)	9.13E-11	130	67,497
509.1	Respiratory failure	5.84	(3.36-10.1)	3.71E-10	82	105,765
418	Nonspecific chest pain	1.98	(1.60-2.45)	4.79E-10	607	86,891
585.1	Acute renal failure	3.16	(2.20-4.55)	5.70E-10	258	104,419
427.9	Palpitations	1.98	(1.59-2.47)	1.65E-09	533	87,904
646	Other complications of pregnancy NEC	5.17	(2.90-9.22)	2.50E-08	52	70,931
292	Neurological disorders	2.60	(1.83-3.68)	7.80E-08	209	108,436
782.3	Edema	2.17	(1.63-2.89)	1.01E-07	355	111,634
350.1	Abnormal involuntary movements	2.54	(1.80-3.60)	1.36E-07	213	114,492
781	Symptoms involving nervous and musculoskeletal systems	3.31	(2.10-5.22)	2.43E-07	123	122,421
1010	Other tests	2.88	(1.91-4.34)	4.48E-07	130	112,904
427.21	Atrial fibrillation	2.30	(1.66-3.19)	5.25E-07	335	87,904
644	Anemia during pregnancy	7.11	(3.22-15.7)	1.22E-06	28	73,007
359.2	Myopathy	11.0	(4.16-29.0)	1.34E-06	21	117,651
599.2	Retention of urine	2.97	(1.88-4.69)	2.90E-06	147	94,396
638	Other high-risk pregnancy	1.99	(1.48-2.69)	6.16E-06	254	121,636
502	Postinflammatory pulmonary fibrosis	6.45	(2.79-14.9)	1.31E-05	33	105,765
649.1	Diabetes or abnormal glucose tolerance complicating pregnancy	4.10	(2.14-7.86)	2.06E-05	43	67,497
420.1	Myocarditis	9.91	(3.42-28.7)	2.37E-05	16	120,148
386.9	Dizziness and giddiness	1.66	(1.31-2.10)	2.85E-05	525	105,771
199	Neoplasm of uncertain behavior	2.91	(1.76-4.80)	3.03E-05	107	116,518
1010.6	Reproductive and maternal health services	1.62	(1.29-2.05)	4.50E-05	416	117,591
1013	Asphyxia and hypoxemia	4.90	(2.26-10.6)	5.42E-05	37	119,659

Sensitivity analysis: propensity-matched cohort.

Phecode	Description	Odds Ratio	95% CI	p value	No. cases	No. controls
512.9	Other dyspnea	2.88	(2.37-3.50)	3.22E-26	487	53,243
512.7	Shortness of breath	2.35	(1.96-2.82)	1.54E-20	602	53,243
278.11	Morbid obesity	2.16	(1.75-2.65)	2.69E-13	402	103,905
427.9	Palpitations	2.15	(1.74-2.66)	1.27E-12	386	94,868
509.1	Respiratory failure	6.44	(3.73-11.1)	2.48E-11	59	106,519
569.2	Gastrointestinal complications	4.54	(2.88-7.16)	8.01E-11	78	110,663
136	Other infectious and parasitic diseases	10.5	(5.04-21.7)	2.74E-10	39	117,892
418	Nonspecific chest pain	1.89	(1.55-2.31)	4.29E-10	446	92,556
418.1	Precordial pain	2.77	(1.98-3.88)	3.34E-09	149	92,556
649	Conditions of the mother complicating pregnancy, childbirth, or the puerperium	2.95	(2.06-4.21)	3.36E-09	126	62,991
585.1	Acute renal failure	2.96	(2.05-4.26)	6.35E-09	149	106,463
427.21	Atrial fibrillation	2.35	(1.73-3.19)	4.17E-08	209	94,868
359.2	Myopathy	15.7	(5.83-42.0)	4.77E-08	26	114,733
782.3	Edema	2.24	(1.67-3.01)	7.19E-08	200	110,984
646	Other complications of pregnancy NEC	4.67	(2.66-8.22)	8.36E-08	52	65,260
1010	Other tests	2.89	(1.93-4.31)	2.24E-07	100	110,868
1013	Asphyxia and hypoxemia	10.9	(4.40-26.9)	2.4E-07	23	115,135
292	Neurological disorders	2.42	(1.71-3.44)	7.63E-07	139	108,388
1010.6	Reproductive and maternal health services	1.59	(1.31-1.93)	3.29E-06	478	113,604
452.2	Deep vein thrombosis [DVT]	3.05	(1.90-4.90)	3.82E-06	73	107,835
350.1	Abnormal involuntary movements	2.23	(1.59-3.13)	3.99E-06	149	110,095
638	Other high-risk pregnancy	1.94	(1.46-2.58)	4.38E-06	215	116,623
278.1	Obesity	1.7	(1.35-2.14)	6.37E-06	348	103,905
599.2	Retention of urine	2.82	(1.78-4.47)	9.77E-06	82	98,114
514	Abnormal findings examination of lungs	2.19	(1.54-3.12)	1.43E-05	155	108,938
502	Postinflammatory pulmonary fibrosis	9.23	(3.34-25.5)	1.8E-05	17	106,519
285	Other anemias	1.85	(1.39-2.46)	2.19E-05	233	101,673
649.1	Diabetes or abnormal glucose tolerance complicating pregnancy	3.61	(1.98-6.56)	2.64E-05	44	62,991
327.32	Obstructive sleep apnea	1.63	(1.30-2.06)	2.64E-05	368	101,184
781	Symptoms involving nervous and musculoskeletal systems	2.58	(1.63-4.10)	5.69E-05	77	117,346

Covariate balance in propensity-matched cohort.

Characteristic	Never Infected	SARS-CoV-2 Positive	Overall	Standardized mean difference[a]
Number in cohort	90,264	30,088	120,352	
Age (median [IQR])	43 [31, 59]	43 [30, 57]	43 [30, 59]	0.004
Sex (%)				0.013
Female	51,747 (57.3)	16,718 (55.6)	68,465 (56.9)	
Male	38,517 (42.7)	13,370 (44.4)	51,887 (43.1)	
Race (%)				
White	60,534 (67.1)	19,176 (63.7)	79,710 (66.2)	0.012
Black	9,654 (10.7)	3,274 (10.9)	12,928 (10.7)	0.027
Other	4,776 (5.3)	1,714 (5.7)	6,490 (5.4)	0.010
Unknown	15,300 (17.0)	5,924 (19.7)	21,224 (17.6)	0.013
Ethnicity (%)				
Non-Hispanic	68,827 (76.3)	21,936 (72.9)	90763 (75.4)	0.010
Hispanic/Latino	3,027 (3.4)	1,217 (4.0)	4244 (3.5)	0.006
Unknown	18,410 (20.4)	69,35 (23.0)	25345 (21.1)	0.013
Received care at VUMC prior to SARS-CoV-2 test (%)	63,028 (69.8)	20,860 (69.3)	83,888 (69.7)	
SARS-CoV-2 testing indication (%)				0.095
Asymptomatic screening	28,982 (32.1)	6,095 (20.3)	35077 (29.1)	
Symptomatic testing	61,282 (67.9)	23,993 (79.7)	85275 (70.9)	
EHR observation time				
Before SARS-CoV-2 test, median [IQR], years	5.6 [0.8, 14.4]	5.8 [0.8, 14.4]	5.7 [0.8, 14.4]	0.001
After recovery, median [IQR], days	393 [242, 509]	361 [285, 427]	381 [252, 500]	0.046
Hospitalization associated with SARS-CoV-2 test (%)	12,434 (13.8)	3,393 (11.3)	15,827 (13.2)	
Time from SARS-CoV-2 test to first follow up (median [IQR])	78 [46, 165]	86 [48, 181]	80 [46, 169]	0.034
Died (%)	531 (0.6)	158 (0.5)	689 (0.6)	0.001
Comorbidites prior to SARS-CoV-2 test (%)				
Myocardial infarction	1,194 (1.3)	419 (1.4)	1,613 (1.3)	0.006
Congestive heart failure	1,811 (2.0)	630 (2.1)	2,441 (2.0)	0.006
Peripheral vascular disease	1,052 (1.2)	325 (1.1)	1,377 (1.1)	0.008
Cerebrovascular disease	1,957 (2.2)	610 (2.0)	2,567 (2.1)	0.01
Dementia	327 (0.4)	137 (0.5)	464 (0.4)	0.015
Chronic pulmonary disease	5,235 (5.8)	1,646 (5.5)	6,881 (5.7)	0.014
Rheumatologic disease	1,555 (1.7)	428 (1.4)	1,983 (1.6)	0.024
Peptic ulcer disease	405 (0.4)	147 (0.5)	552 (0.5)	0.006
Diabetes	4,993 (5.5)	1,964 (6.5)	6,957 (5.8)	0.042
Mild liver disease	2,195 (2.4)	726 (2.4)	2,921 (2.4)	0.001
Severe liver disease	528 (0.6)	170 (0.6)	698 (0.6)	0.003
Hemiplegia or paraplegia	458 (0.5)	153 (0.5)	611 (0.5)	0.001
Renal disease	1,627 (1.8)	750 (2.5)	2,377 (2.0)	0.048
Any malignancy	5,268 (5.8)	1,463 (4.9)	6,731 (5.6)	0.043
Metastatic solid tumor	824 (0.9)	246 (0.8)	1,070 (0.9)	0.01
AIDS or HIV infection	606 (0.7)	173 (0.6)	779 (0.6)	0.012

Variables used to develop the propensity-scoring model for probability of testing positive for SARS-CoV-2 included age, sex, race, ethnicity, symptomatic testing indication, inpatient hospitalization around time of SARS-CoV-2 test, observation time after recovery, and length of EHR prior to SARS-CoV-2 testing.

[a] Standardized mean difference values of less than 0.1 indicate acceptable matching between groups.