Journal of the American Medical Informatics Association, 00(0), 2022, 1–12
https://doi.org/10.1093/jamia/ocac159
Research and Applications

Research and Applications

Scanning the medical phenome to identify new diagnoses after recovery from COVID-19 in a US cohort

Vern Eric Kerchberger[1,2], Josh F. Peterson[1,2], and Wei-Qi Wei[1]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee, USA and [2]Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, USA

Corresponding Author: Vern Eric Kerchberger, MD, Division of Allergy, Pulmonary, and Critical Care Medicine, Department of Medicine, Vanderbilt University Medical Center, 1161 21st Ave S, Ste T-1218, Nashville, TN 37232, USA; vern.e.kerchberger@vumc.org

Received 7 April 2022; Revised 29 June 2022; Editorial Decision 9 August 2022; Accepted 23 August 2022

ABSTRACT

Objective: COVID-19 survivors are at risk for long-term health effects, but assessing the sequelae of COVID-19 at large scales is challenging. High-throughput methods to efficiently identify new medical problems arising after acute medical events using the electronic health record (EHR) could improve surveillance for long-term consequences of acute medical problems like COVID-19.
Materials and Methods: We augmented an existing high-throughput phenotyping method (PheWAS) to identify new diagnoses occurring after an acute temporal event in the EHR. We then used the temporal-informed phenotypes to assess development of new medical problems among COVID-19 survivors enrolled in an EHR cohort of adults tested for COVID-19 at Vanderbilt University Medical Center.
Results: The study cohort included 186 105 adults tested for COVID-19 from March 5, 2020 to November 1, 2021; of which 30 088 (16.2%) tested positive. Median follow-up after testing was 412 days (IQR 274–528). Our temporal-informed phenotyping was able to distinguish phenotype chapters based on chronicity of their constituent diagnoses. PheWAS with temporal-informed phenotypes identified increased risk for 43 diagnoses among COVID-19 survivors during outpatient follow-up, including multiple new respiratory, cardiovascular, neurological, and pregnancy-related conditions. Findings were robust to sensitivity analyses, and several phenotypic associations were supported by changes in outpatient vital signs or laboratory tests from the pretesting to postrecovery period.
Conclusion: Temporal-informed PheWAS identified new diagnoses affecting multiple organ systems among COVID-19 survivors. These findings can inform future efforts to enable longitudinal health surveillance for survivors of COVID-19 and other acute medical conditions using the EHR.

Key words: COVID-19, COVID-19/complications, electronic health records, cohort study, phenome-wide association study

INTRODUCTION

The coronavirus disease 2019 (COVID-19) pandemic continues to evolve, with more than 400 million confirmed cases worldwide over numerous waves.[1] Although most COVID-19 patients ultimately recover, many survivors report new medical problems arising after recovery from their acute illness.[2–15] 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,[5,9,11,15] however, these studies relied upon proprietary commercial claims or administrative data,[9] unique national databases,[5,11] or employed complex feature engineering and advanced statistical methods,[11,15] 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 (EHRs).[16–20] PheWAS has a well-documented R package incorporating feature engineering and analysis methods to facilitate study design and harmonization of research.[17,18,21] There also is increasing use of PheWAS to investigate the phenotypic consequences of nongenetic variables such as race, healthcare costs, or comorbidity burden.[22–29] 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.

Objective

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.[30,31] 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 Supplementary Appendix.

Defining postacute 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 postacute COVID-19 as 4 weeks after onset of symptoms,[2,3,11] we defined recovery from acute disease and transition to the postacute phase as either 30 days after SARS-CoV-2 testing for nonhospitalized 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 postacute 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.[18,20,32] We also collected vital sign values and results of common clinical laboratory tests obtained both prior to SARS-CoV-2 testing and after the postacute 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 postacute period. In keeping with usual practice for PheWAS, we defined "phenotype cases" as patients with a corresponding phecode on at least 2 separate days, and "phenotype controls" as patients with zero codes.[18,21] 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".[23,32]

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 2 datasets based on occurrence of the diagnosis code relative to the event of interest (in this study, transition to the postacute phase, Figure 1). We applied the PheWAS feature engineering method to the pre-event and postevent 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 postevent 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 (Supplementary Table S1 and Appendix).

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:

$$p_{\text{retention}} = \frac{\eta_{\text{temporal-informed}}}{\eta_{\text{naive}}} \quad (1)$$

Where $n_{\text{temporal-informed}}$ is the phenotype case or control counts using temporal-informed phenotyping and n_{naive} is the phenotype

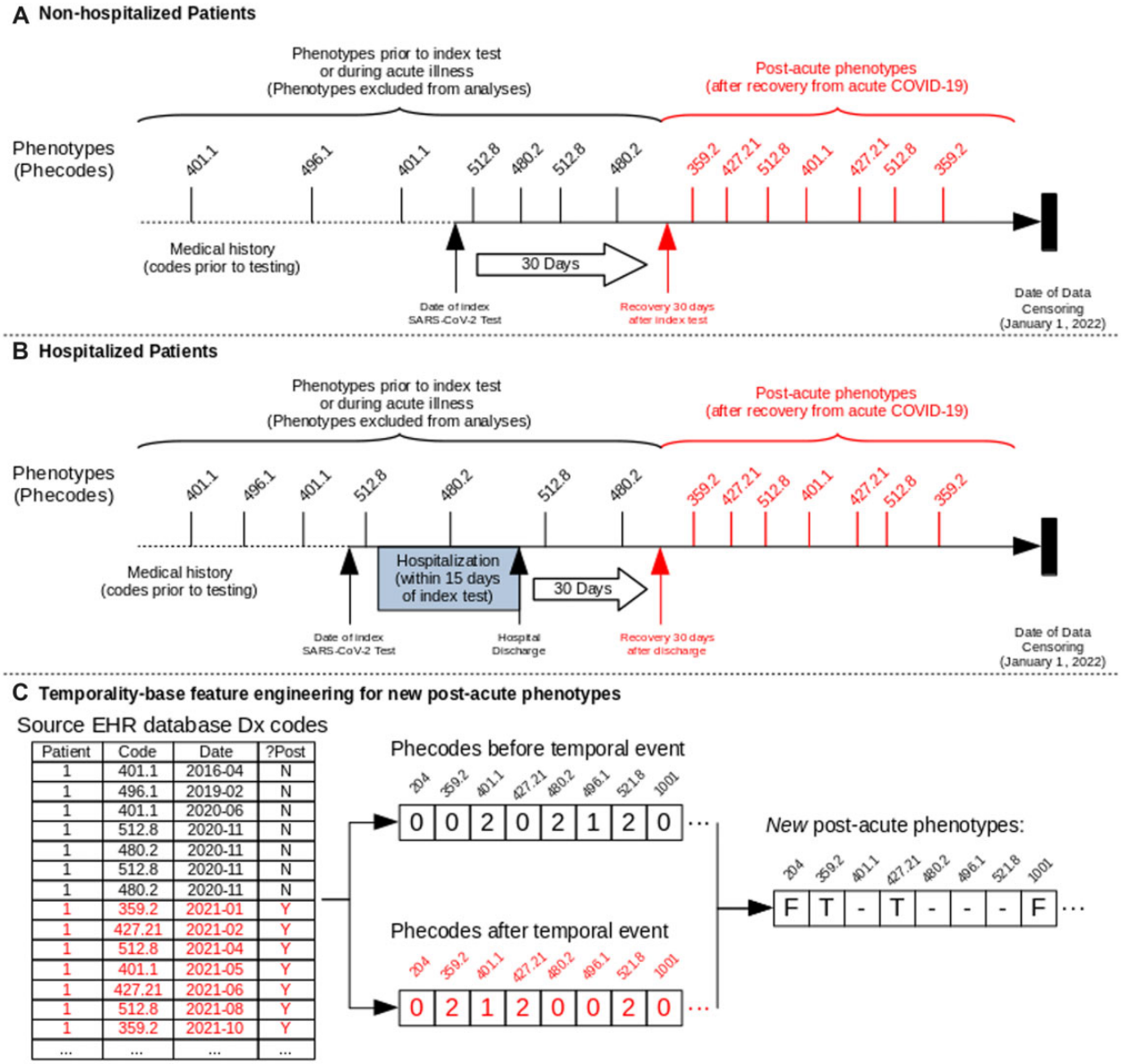

Figure 1. Graphical timeline of data collection from electronic health record and phenotype encoding schematic. Graphical timeline of index SARS-CoV-2 test, recovery, and phenotype case and control definitions for (A) patients who were not hospitalized or (B) were hospitalized around time of index SARS-CoV-2 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 nonhospitalized 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 "postevent" 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 2 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 postevent datasets (eg, 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 (Supplementary Table S1 and Appendix).

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.[33–35] We performed PheWAS using

logistic regression to model the log-odds of developing each temporal-informed phenotype in the postacute period given the presence or absence of the exposure of interest, adjusting for demographic and comorbidity covariates as:

$$\text{logit } p(Y_i = 1 \mid \text{Exposure, Covariates}) = \beta_0 + \beta_{\text{EXPOSURE}} \times \text{Exposure} + \beta_{\text{COVAR}} \times \text{Covariates} \quad (2)$$

where $i=\{1, \ldots, n\}$ phecodes with at least 10 phenotype cases in the cohort.[20,32] For vital signs and clinical laboratory tests, we modeled the change in value from pretesting to the postacute period as:

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

where $Y_{\text{pre-testing}}$ is the median value from all outpatient measurements obtained within 180 days prior to SARS-CoV-2 testing and $Y_{\text{postacute}}$ is the median value from all outpatient measurements within 365 days after entering the postacute phase. Comorbidities were ascertained using a phecode-based mapping of the Charlson comorbidities (Supplementary Table S2 and Appendix).[36] 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% CIs. Statistical significance was set using a Bonferroni correction for number of independent tests. Additional details on model covariates and sensitivity analyses are provided in Supplementary Appendix. All analyses were performed using the R package *PheWAS*.[21]

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).[37,38] 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 9755 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 postacute phase, leaving 186 105 adults in the primary cohort (Supplementary Figure S1 and Appendix). Among these, 30 088 (16.2%) tested positive. Median age at initial test was 46 years (IQR 32–61), 57.1% were female, and 4677 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 1 follow-up visit in our system after recovery. Additional demographic and clinical characteristics of the study population are shown in Table 1 and Supplementary Table S3 (Supplementary Appendix).

Temporal-informed phenotyping of postacute period

At the data censoring date and after mapping for diagnosis-specific exclusions, 1347 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 postacute 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 2 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 postacute period, whereas neoplasms were least likely to represent new diagnoses (Supplementary Table S4 and Appendix). Control retention under temporal-informed phenotyping was high (per-phecode median 91.7%; IQR: 87.9%–95.1%; Supplementary Figure S2 and Appendix), although several respiratory phenotypes (eg, 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 (Supplementary Figure S3 and Appendix). Patients with ≥6 months of care at VUMC prior to testing were more likely to have at least 1 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%, $P < 1.0 \times 10^{-15}$), indicating the temporal-informed phenotypes were not driven by patients with short EHR histories.

Temporal-informed PheWAS identifies new postacute 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 3, Table 2). 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. In contrast, the naive approach identified 219 phenotypes reaching Bonferroni-adjusted significance (Supplementary Table S5, Figure S4, and Appendix). 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 (eg, altered mental status, hypotension, respiratory failure, sepsis, septicemia, acidosis) or chronic medical conditions know to be risk factors for COVID-19 (eg, chronic kidney disease, essential hypertension, hyperlipidemia).[24,25] 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 postinflammatory pulmonary fibrosis were only

Table 1. 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)	3274 (10.9)	20 380 (11.0)
Other race or multiracial	7901 (5.1)	1714 (5.7)	9615 (5.2)
Unknown/not reported	18 996 (12.2)	5924 (19.7)	24 920 (13.4)
White	112 014 (71.8)	19 176 (63.7)	131 190 (70.5)
Ethnicity (%)			
Hispanic/Latino	4759 (3.1)	1217 (4.0)	5976 (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)	6935 (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)	6095 (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)	3393 (11.3)	46 539 (25.0)
Severe COVID-19 (%)[d]	–	2358 (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/anticoagulation visit	42 646 (27.3)	7216 (24.0)	49 862 (26.8)
Inpatient surgery or procedure	27 213 (17.4)	4091 (13.6)	31 304 (16.8)
Telemedicine visit	16 617 (10.7)	2478 (8.2)	19 095 (10.3)
Outpatient surgery or procedure	19 725 (12.6)	2728 (9.1)	22 453 (12.1)
Allied health practitioner visit[f]	14 821 (9.5)	2580 (8.6)	17 401 (9.4)
Infusion/radiation care	4043 (2.6)	542 (1.8)	4585 (2.5)
Maternity care	3899 (2.5)	482 (1.6)	4381 (2.4)
Outpatient observation in Emergency Department	2403 (1.5)	422 (1.4)	2825 (1.5)
Inpatient medical admission	1197 (0.8)	1239 (4.1)	2436 (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 (%)	7565 (4.8)	609 (2.0)	8174 (4.4)
Pregnant around time of SARS-CoV-2 test (%)	4488 (2.9)	189 (0.6)	4677 (2.5)
Died during postacute phase (%)	1535 (1.0)	158 (0.5)	1693 (0.9)

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

[b]Reasons for asymptomatic screening included: asymptomatic admission to the hospital for another diagnosis, preprocedural or presurgical screening, known SARS-CoV-2 exposure, prereceipt of immunosupressive or antineoplastic therapy, pretransplant evaluation, or requirement for placement in postacute 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 1 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.

seen using temporal-informed analyses. Strength of associations (based on *P* value) was higher under the naive approach due to higher phenotype case counts, but adjusted odds ratios were similar under both approaches (Supplementary Figure S5 and Appendix).

Figure 4 illustrates subgroup analyses based on demographics and timing of the postacute diagnoses. New postacute 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 postacute period, however, 14 phenotypes presented earlier during the first 60 days after recovery. Subgroup PheWAS results are available in the Supplementary Appendix (Supplementary Tables S6–S8). Our findings were also robust to 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

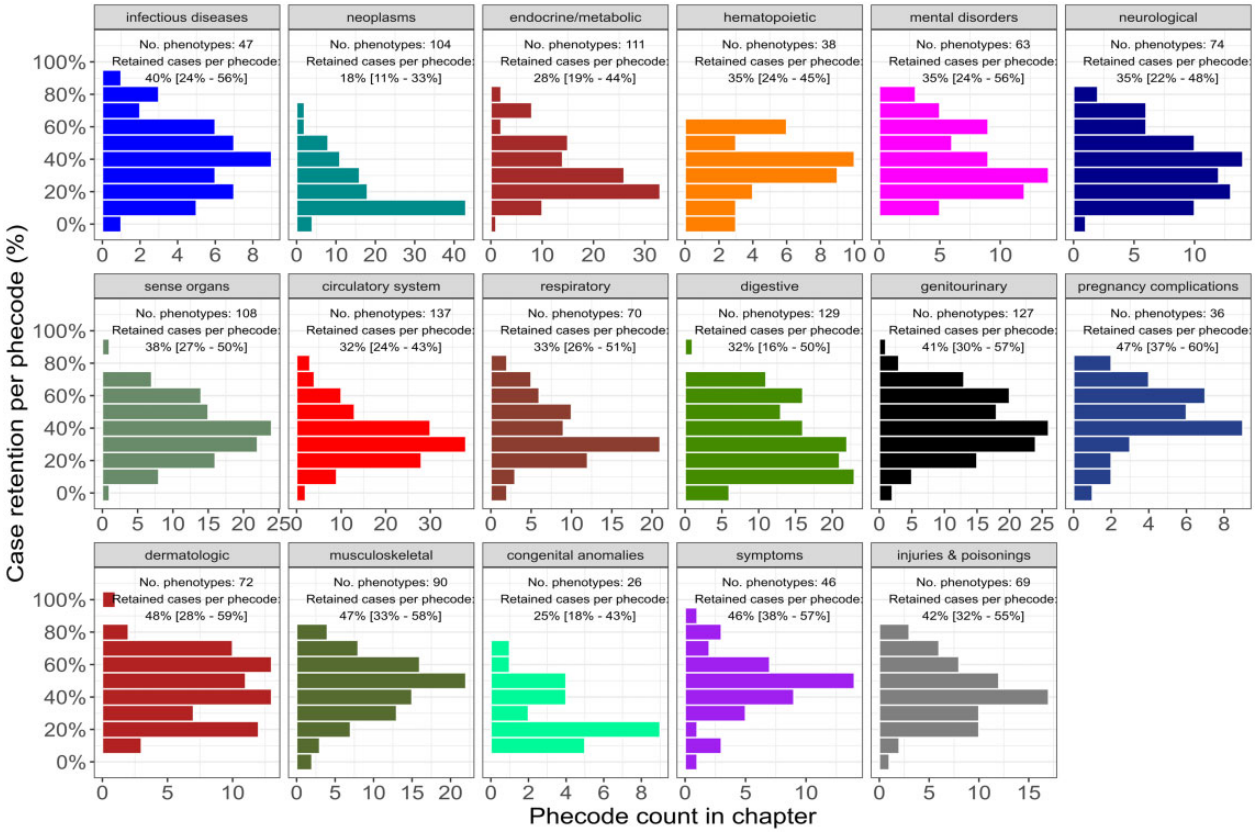

Figure 2. Phecode case retention by temporal-informed phenotyping. Histograms of phenotype case retention per PheWAS code (phecode) using temporal-informed phenotyping. Individual histograms indicate each chapter within the phecode hierarchy.[32] Number of phecodes per chapter are shown on *x* axis, 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.

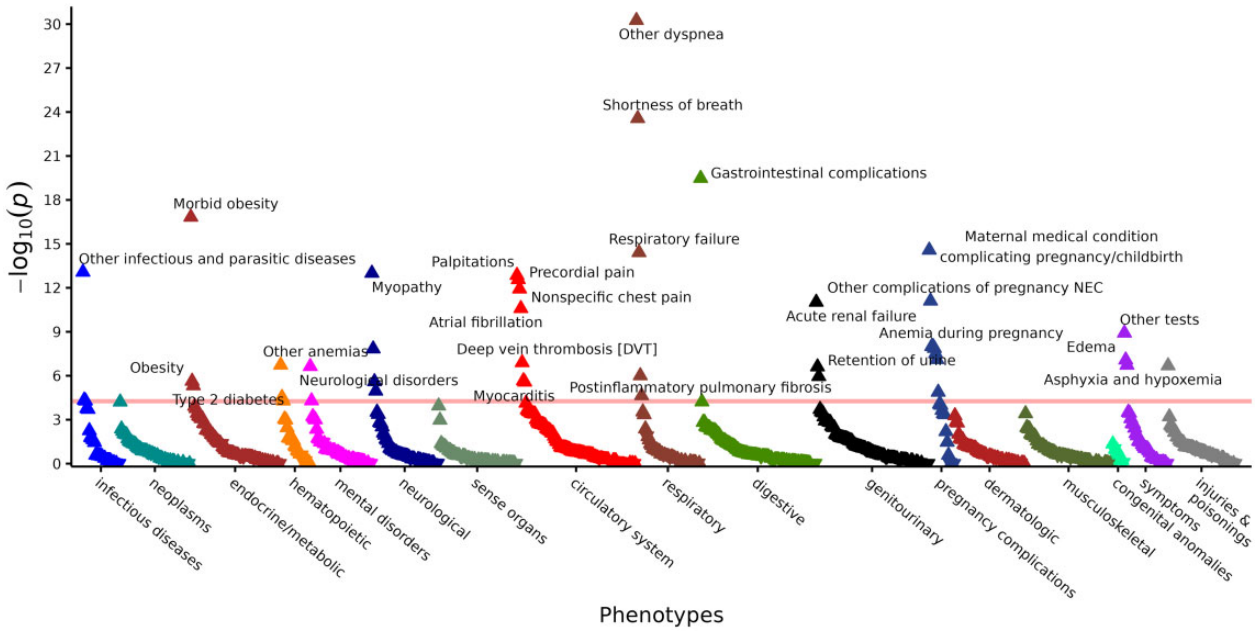

Figure 3. Temporal-informed phenome scan of postacute COVID-19. PheWAS plot of new postacute phenotypes identified by temporal-informed phenotyping for COVID-19 survivors vs never-infected patients as the referent group (n=186 105, phenotypes available for testing=902). The *x* axis represents phecodes grouped by chapter within the phecode hierarchy.[32] 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 2. Summary of temporal-informed PheWAS in postacute 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}	1698	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	Postinflammatory 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.[32]

controls to each COVID-19 survivor (Supplementary Tables S9–S12, Figure S6, and Appendix).

Postacute clinical phenotypes associated with severe COVID-19

Among the 30 088 COVID-19 survivors, those with severe disease ($n = 2358$, 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 postacute phenotypes associated with severe SARS-CoV-2 survivors are shown in Table 3.

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 (eg, 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 pretesting to postacute 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 pretesting and postacute BMI recorded in the EHR,

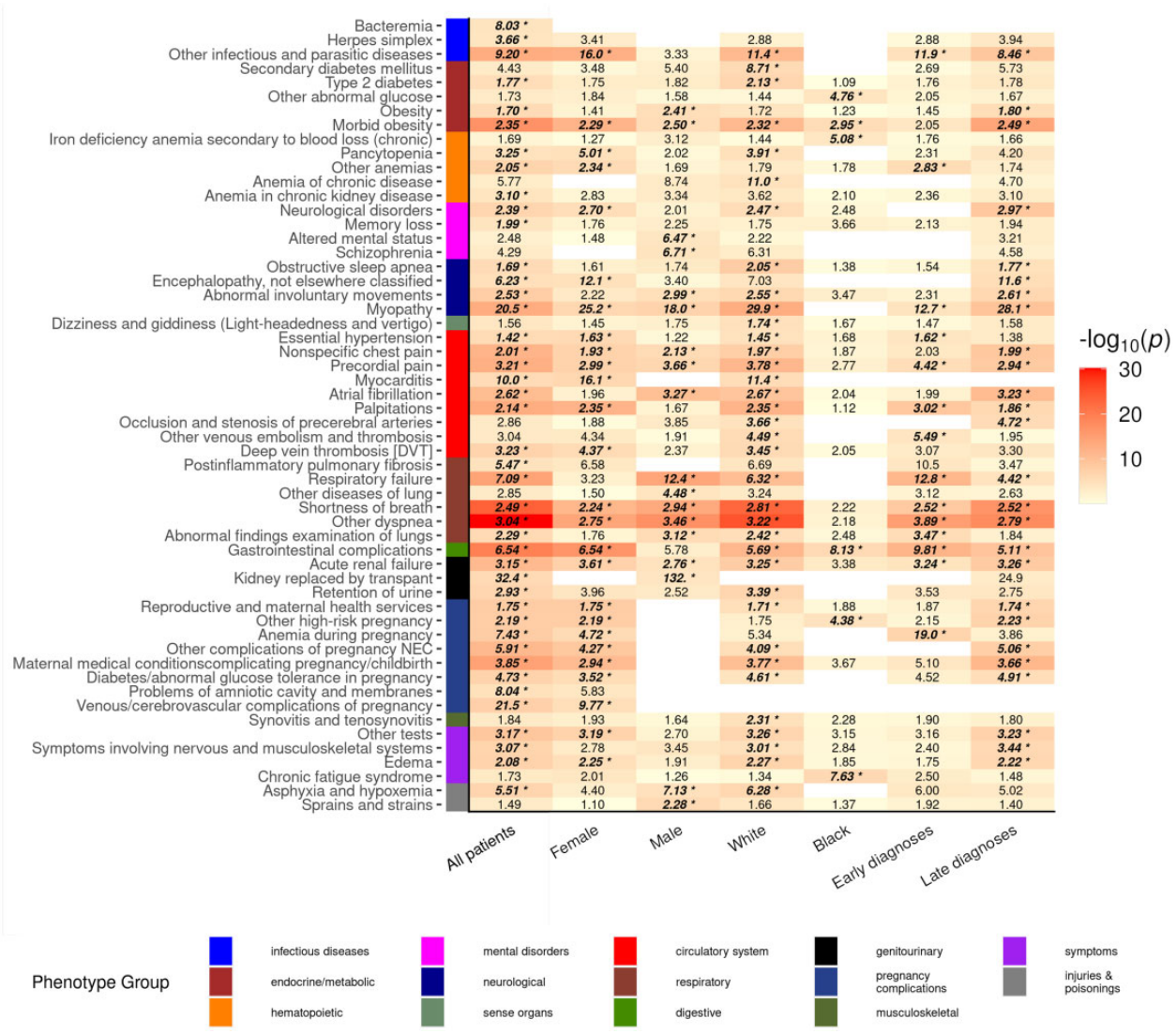

Figure 4. Temporal-informed phenome scans of postacute COVID-19 by demographic subgroups and timing of postacute diagnoses. PheWAS results for new postacute phenotypes identified by temporal-informed phenotyping among all adults tested for SARS-CoV-2 (left column, n=186 105), stratified by demographic 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.[32] 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 (<10) to perform the analysis for that phenotype in the subgroup.

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; $P = 2.00 \times 10^{-13}$). 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 4, Figure 5). 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 (Supplementary Figure S7).

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 postacute 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.[2,3] Several postacute phenotype associations were also supported by changes in vital signs values from pretesting to the postacute period. Although the observed differences in

Table 3. 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	Postinflammatory 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]

Table 4. Changes in outpatient vital signs or laboratory studies for select temporal-informed phenotypes

Postacute phenotype(s)	Vital sign/lab (units)	Subgroup[b]	Change in lab or vital sign from pretesting to postacute[a]			
			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)	Nonobese ($n = 37\,838$)	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 prehypertension ($n = 28\,912$)	−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 ($10^3/\mu L$)	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-CoV-2 testing and within 365 days following recovery.

[b]Prior to SARS-CoV-2 testing.

[c]Calculated for each patient as $Y_{postacute} - Y_{pretesting}$, where Y is the vital sign value or laboratory value. Negative values indicate a decrease in the vital sign/lab value from the pretesting to the postacute phases, and positive values indicate an increase in the vital sign/lab value.

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

[e]Adjusted *P* values using linear regression.

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.[39] 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.[2,4–14] Ayoubkhani et al[5] 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. Daugherty et al[9] 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. Al-Aly et al[11] 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. Similar to our findings of increased myopathy, neurological deficits, encephalopathy, and memory loss, Taquet et al[12] found that COVID-19 survivors had elevated risk for developing multiple neurologic and psychiatric disorders in a multinational EHR dataset. Estiri et al[15] evaluated the temporal evolution postacute 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.

Strengths

Our temporal-informed phenotyping framework naturally augments classical PheWAS, allowing us to identify potential postacute sequelae of COVID-19 and replicate several associations identified in other studies. The distribution of case retention under

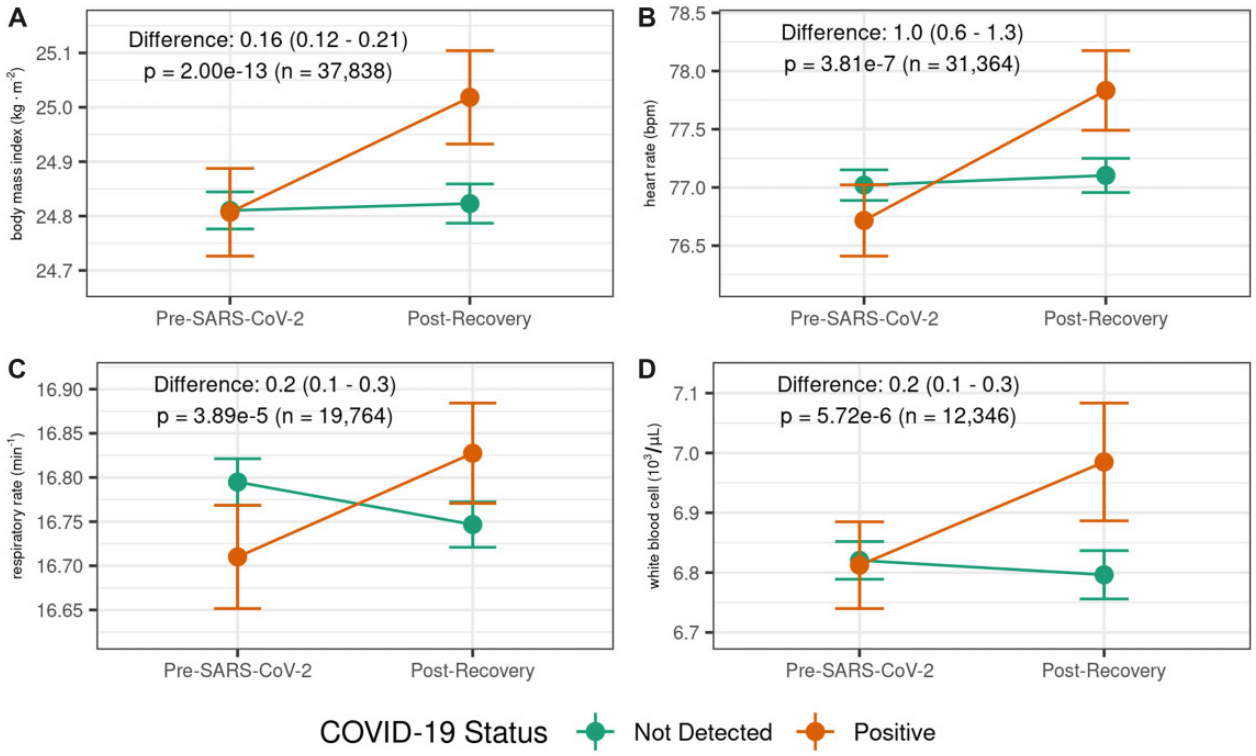

Figure 5. Changes in select vital signs and laboratory test values in postacute COVID-19. COVID-19 survivors had more substantial changes in (A) body mass index, (B) heart rate, (C) respiratory rate, and (D) white blood cell count from pretesting to postrecovery compared with never-infected controls. For each patient we used the median pretesting values obtained during outpatient visits occurring within 180 days before the index SARS-CoV-2 test, and the median postrecovery 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.

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.[15,22,40–43] 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 ~68 000 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.[19–21] PheWAS analyses are also more accessible to researchers than more complex machine learning methods.[44] Thus, our temporal-informed phenotyping could be easily adapted to examine the postacute phenotype consequences among survivors other acute medical event such as pneumonia or sepsis.[45,46]

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 1 follow-up visit in the postacute phase. This study leveraged our longstanding institutional experience with using the EHR for secondary research,[19,30] allowing us to capture deep phenotyping information, such as SARS-CoV-2 testing indication and setting of postacute diagnoses, which may not be well-represented in administrative datasets or cross-institutional research databases.[11,47] We were also able to compare temporal-informed phenotypes between survivors of severe COVID-19 vs survivors of nonsevere COVID-19, and we correlated several temporal-informed phenotypic associations with changes in vital signs or laboratory values from the pretesting to postacute 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 (eg, 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,[31,48] 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 postacute 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.[43] 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.[49] 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.

FUNDING

This study was supported in part by the National Institutes of Health continuing education grant NIH T15 LM007450 (VEK); research grants NIH K01 HL157755-01 (VEK), NIH U01 HG01166-01S1 (JFP), and NIH R01 GM139891-01 (W-QW); the American Thoracic Society (VEK), and the Francis Family Foundation (VEK). The project described was also supported by CTSA award No. UL1 TR002243 from the National Center for Advancing Translational Sciences. Its contents are solely the responsibility of the authors and do not necessarily represent official views of the National Center for Advancing Translational Sciences or the National Institutes of Health.

AUTHOR CONTRIBUTIONS

JFP and WQW led development and design of the institutional registry used for the study. VEK and WQW were responsible for study conceptualization and design, and verified accuracy and integrity of the data. VEK acquired the study data, performed the analyses and data visualizations, and wrote the first draft of the manuscript. WQW provided computing resources for execution of the study and supervised the analyses. All authors contributed to interpretation of the results, revised the manuscript critically for intellectual content, and approved the final manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

Individual-level data used in this study cannot be made publicly available due to institutional controls on patient health information used in secondary research. Requests for a deidentified dataset with data dictionary may be made to the corresponding author.

REFERENCES

1. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis* 2020; 20 (5): 533–4.
2. Nalbandian A, Sehgal K, Gupta A, *et al*. Post-acute COVID-19 syndrome. *Nat Med* 2021; 27 (4): 601–15.
3. 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 (22): 2251–2.
4. Logue JK, Franko NM, McCulloch DJ, *et al*. Sequelae in adults at 6 months after COVID-19 infection. *JAMA Netw Open* 2021; 4 (2): e210830.
5. 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.
6. The Writing Committee for the COMEBAC Study Group. Four-month clinical status of a cohort of patients after hospitalization for COVID-19. *JAMA* 2021; 325: 1525–34.
7. Sonnweber T, Sahanic S, Pizzini A, *et al*. Cardiopulmonary recovery after COVID-19: an observational prospective multicentre trial. *Eur Respir J* 2021; 57 (4): 2003481.
8. 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 (4): 399–401.
9. 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.
10. 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 (6): 892–6.
11. Al-Aly Z, Xie Y, Bowe B. High-dimensional characterization of post-acute sequelae of COVID-19. *Nature* 2021; 594 (7862): 259–64.
12. 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 (5): 416–27.
13. 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: 101019.
14. Huang L, Yao Q, Gu X, *et al*. 1-Year outcomes in hospital survivors with COVID-19: a longitudinal cohort study. *Lancet* 2021; 398 (10302): 747–58.

15. Estiri H, Strasser ZH, Brat GA, *et al.*; Consortium for Characterization of COVID-19 by EHR (4CE). Evolving phenotypes of non-hospitalized patients that indicate long COVID. *BMC Med* 2021; 19 (1): 249.
16. 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 (9): 1205–10.
17. 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 (1): e1003087.
18. 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 (12): 1102–11.
19. 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.
20. 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 (7): e0175508.
21. 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 (16): 2375–6.
22. 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 (e2): e281–7.
23. Bastarache L. Using phecodes for research with the electronic health record: from PheWAS to PheRS. *Annu Rev Biomed Data Sci* 2021; 4: 1–19.
24. 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 (11): e0242182.
25. 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 (7): 1351.
26. Zhang T, Goodman M, Zhu F, *et al.* Phenome-wide examination of comorbidity burden and multiple sclerosis disease severity. *Neurol Neuroimmunol Neuroinflamm* 2020; 7 (6): e864.
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 (3): 760–7.
28. Dashti HS, Cade BE, Stutaite G, *et al.* Sleep health, diseases, and pain syndromes: findings from an electronic health record BioBank. *Sleep* 2021; 44 (3): zsaa189.
29. 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 (2): 187–94.
30. Danciu I, Cowan JD, Basford M, *et al.* Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform* 2014; 52: 28–35.
31. 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.
32. 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 (4): e14325.
33. 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 (11): 1061–9.
34. 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 (5): 475–81.
35. Huang C, Huang L, Wang Y, *et al.* 6-Month consequences of COVID-19 in patients discharged from hospital: a cohort study. *Lancet* 2021; 397 (10270): 220–32.
36. 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 (1): e187223.
37. Benchimol EI, Smeeth L, Guttmann A, *et al.*; RECORD Working Committee. The reporting of studies conducted using observational routinely-collected health data (RECORD) statement. *PLoS Med* 2015; 12 (10): e1001885.
38. 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.
39. 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 (3): E53–63.
40. 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.
41. 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.
42. 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.
43. 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.
44. Pfaff ER, Girvin AT, Bennett TD, *et al.*; N3C Consortium. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit Health* 2022; 4 (7): e532–41.
45. 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 (9): 1065–74.
46. Corrales-Medina VF, Alvarez KN, Weissfeld LA, *et al.* Association between hospitalization for pneumonia and subsequent risk of cardiovascular disease. *JAMA* 2015; 313 (3): 264–74.
47. Haendel MA, Chute CG, Bennett TD, *et al.* The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc* 2021; 28 (3): 427–43.
48. 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 (8): 2532–5.
49. Hodcroft EB. CoVariants: SARS-CoV-2 mutations and variants of interest. 2021. https://covariants.org/. Accessed August 18, 2021.