JAMA Network Open

Original Investigation | Infectious Diseases

Long COVID Persistence and Surveillance Gaps Across 58 US Hospitals

Jiazi Tian, MSc; Alaleh Azhir, MD, MSc; Matthew Decaro, MSc; Ngan Chau, BS; Jonas Hügel, PhD; Michele Morris, BA; Jingya Cheng, MB; Pedram Fard, PhD; Ingrid V. Bassett, MD, MPH; Douglas S. Bell, MD, PhD; Elmer V. Bernstam, MD, MSE; Shyam Visweswaran, MD, PhD; Jeffrey G. Klann, PhD; Shawn N. Murphy, MD, PhD; Hossein Estiri, PhD

Abstract

IMPORTANCE Surveillance of postacute sequelae of SARS-CoV-2 infection (PASC) depends on diagnostic coding systems that capture fewer than one-half of affected individuals, rendering millions invisible to health systems and policymakers.

OBJECTIVE To quantify the gap between true PASC burden and diagnostic code–based estimates, determine the proportion representing chronic disease, and characterize organ system heterogeneity and temporal trends across diverse populations.

DESIGN, SETTING, AND PARTICIPANTS This retrospective cohort study used electronic health record data from 58 hospitals and affiliated clinics in 4 US regions, from 2017 to 2025. Adults (aged ≥18 years) with laboratory-confirmed SARS-CoV-2 infection or a COVID-19 diagnosis code were included. A custom artificial intelligence algorithm, the Precision Phenotyping for Research Cohorts (P2RC), was implemented using federated infrastructure.

EXPOSURE Laboratory-confirmed SARS-CoV-2 infection or COVID-19 diagnosis code.

MAIN OUTCOMES AND MEASURES The primary outcomes were PASC prevalence, the proportion classified as chronic conditions, organ system distribution, and temporal trends from 2020 to 2024. χ^2 Tests were used to assess organ system heterogeneity across regions, and negative binomial regression was used to model quarterly temporal trends, yielding incidence rate ratios (IRRs) with 95% CIs.

RESULTS In this cohort study of 457 950 COVID-19 cases (mean age, 52.05 years; 275 107 [60.07%] female), the P2RC algorithm identified 74 560 PASC cases (16.28% overall; 28 585 [18.58%] in New England, 978 [19.55%] in Southeast Texas, 10 534 [22.69%] in Southern California, and 34 463 [13.64%] in Western Pennsylvania), more than 2-fold higher than the proportion identified by code-based surveillance (<7%). Of 883 *International Statistical Classification of Diseases, Tenth Revision, Clinical Modification* codes associated with PASC, 594 (67.27%) represented chronic or potentially chronic conditions. Of 74 560 patients with PASC, 66 587 (89.31%) developed chronic conditions requiring ongoing clinical management; this represents 14.54% of the total number of 457 950 patients with COVID-19. Substantial organ system heterogeneity was observed (χ^2 = 2504.73; P < .001): New England demonstrated thyroid-predominant endocrine patterns, while Southeast Texas, Southern California, and Western Pennsylvania showed metabolic-predominant profiles. Negative binomial regression revealed increasing PASC prevalence through mid-2024 (IRR per quarter, 1.01 [95% CI, 1.00-1.01; P < .001] in New England; 1.00 [95% CI, 1.00-1.01; P < .001] in Southern California; and 1.02 [95% CI, 1.01-1.02; P < .001] in Western Pennsylvania), indicating an accumulating rather than resolving burden.

(continued)

Key Points

Question What is the true burden of chronic disease following COVID-19, and why does current surveillance fail to capture it?

Findings In this cohort study of 457 950 patients with COVID-19 across 58 hospitals, validated computable phenotyping identified postacute sequelae of SARS-CoV-2 infection in 16.28% of cases, 2-fold higher than diagnostic code–based surveillance. Of identified manifestations, 89.31% represented chronic conditions, with prevalence increasing through mid-2024.

Meaning These findings suggest that approximately 1 in 6 patients with COVID-19 develops postacute sequelae, predominantly chronic conditions currently invisible to surveillance systems, representing an accumulating rather than resolving health care burden.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

CONCLUSIONS AND RELEVANCE In this cohort study, approximately 1 in 6 patients with COVID-19 developed PASC, and 89.31% of these patients had at least 1 chronic condition. Current diagnostic coding captured fewer than one-half of the cases, obscuring a substantial chronic disease burden. The persistently increasing prevalence through 2024 indicated an accumulating health care burden requiring investment in surveillance infrastructure and integrated care pathways.

JAMA Network Open. 2026;9(5):e2614909. doi:10.1001/jamanetworkopen.2026.14909

Introduction

Large-scale, clinical practice data platforms have enabled population-level research on postacute sequelae of SARS-CoV-2 infection (PASC), yet administrative claims data harbor inherent limitations, including standardization losses during cross-institutional aggregation,[1-4] demographic missingness,[5-7] and systematic ascertainment bias.[8,9] A fundamental challenge confronting PASC surveillance is the marked discrepancy between prevalence estimates derived from clinical assessments and those from administrative coding.[10-12] Meta-analyses estimate PASC prevalence at 43%,[13,14] whereas diagnostic code–based analyses report substantially lower figures. The U09.9 code demonstrates particularly poor sensitivity (4.9%-19.0% across health systems),[15] with population-level analyses finding it assigned to fewer than 1% of COVID-19 survivors.[16,17] Given that U09.9 underascertainment is well established, the critical question shifts from whether diagnostic coding underestimates PASC to the true burden and the proportion that represents chronic disease requiring sustained clinical management.

The public health implications are substantial. Patients with PASC demonstrate elevated health care utilization, including increased specialist consultations and higher hospitalization rates (odds ratio, 1.28; 95% CI, 1.23-1.33).[5] More critically, longitudinal studies reveal that PASC predominantly manifests as chronic rather than self-limited illness. Systematic reviews demonstrate that 45% of COVID-19 survivors experience unresolved symptoms at approximately 4 months,[18] with 30% reporting persistent symptoms at 24 months.[19] This chronicity profile signals an emerging chronic disease epidemic requiring sustained clinical management and health care infrastructure investment.

The Precision Phenotyping for Research Cohorts (P2RC) algorithm is a custom artificial intelligence (AI) system that operationalizes World Health Organization case definitions through transitive Sequential Pattern Mining of temporal electronic health record (EHR) sequences, achieving 80% precision with a prevalence estimate of 23%.[20] Emerging evidence suggests PASC comprises multiple distinct endotypes characterized by predominant organ system involvement,[12,19] although whether these patterns vary systematically across health care settings and populations remains inadequately characterized. We conducted multisite implementation of the P2RC algorithm across 4 geographically and demographically diverse US regions to address 4 objectives: quantify true PASC prevalence vs diagnostic code–based estimates, determine the proportion representing chronic disease burden, delineate organ system heterogeneity patterns, and assess temporal trends spanning 2020 to 2024 to distinguish accumulating vs resolving disease patterns.

Methods

Study Design and Setting

We conducted a retrospective, multicenter cohort study using EHR data from 4 US regions in the Evolve to Next-Gen ACT (ENACT) network,[21] including 12 hospitals in New England, 1 hospital in Southeast Texas, 5 hospitals in Southern California, and 40 hospitals in Western Pennsylvania. All sites included records from affiliated community health centers, outpatient clinics, and telemedicine encounters.

This study was approved by the institutional review boards at participating institutions, with a waiver of informed consent granted due to its retrospective design and the use of deidentified data. This study followed the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guidelines for cohort studies.

Study cohorts were defined using the ENACT cohort discovery platform, which provides a harmonized ontology within the information for integrating biology and the bedside (i2b2) clinical research framework.[22,23] Data were mapped to the Clinical Classifications Software Refined (CCSR),[24] aggregating *International Statistical Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM)* and Procedure Coding System codes into clinically meaningful categories. The ENACT loyalty score[25] was used to select patients with sufficient data continuity. Only adults (aged ≥18 years) were included. Race and ethnicity data were extracted from the EHR at each participating institution, where classifications are derived from patient self-report at the time of registration (ie, Asian, Black, Hispanic, White, unknown, or other, which includes American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Middle Eastern or North African, multiple races, and other unspecified groups). Race and ethnicity are included in this study to characterize the demographic diversity of the multisite cohort.

PASC Phenotype Definition

PASC cases were identified using the P2RC algorithm,[20] which operationalizes the World Health Organization case definition as a diagnosis of exclusion. COVID-19 was ascertained by a documented positive PCR test or an *ICD-10-CM* diagnosis code. Records within 90 days of the first positive test were grouped as a single infection episode. The algorithm uses transitive sequential pattern mining[26] to longitudinal EHR data to identify temporal symptom patterns occurring 3 or more months after infection and persisting for 2 months or longer, while accounting for alternative diagnostic explanations, through an attention mechanism that excludes sequelae explained by preexisting conditions while including chronic conditions temporally associated with COVID-19. Validation demonstrated 79.9% precision in identifying PASC cases.[20] Open-source R code and implementation guidelines are publicly available.[27]

To improve case identification, we applied a clinical exclusion step to remove patients with preexisting conditions unlikely to represent true PASC, thereby enhancing the specificity while maintaining the algorithm's sensitivity for detecting genuine postacute manifestations. Among confirmed PASC cases, a clinical expert (A.A.) classified each *ICD-10-CM* code into 4 chronicity categories: definitively chronic, can be chronic, acute, and variable. Details of the classification system and final *ICD-10-CM* codes are available in eTable 1 in Supplement 1.

We did not separately quantify the prevalence of U09.9 codes at participating sites, as multiple prior studies have established sensitivity below 20% across diverse health systems.[15-17] Our objective was to estimate the true PASC burden through validated phenotyping, rather than redocumenting the known limitations of diagnostic coding.

Multisite Implementation

Each region used the ENACT i2b2 framework to identify patients with laboratory-confirmed SARS-CoV-2 infection or COVID-19 diagnosis codes (eFigure 1 in Supplement 2). Diagnostic and procedure codes were mapped to CCSR categories, and standardized variables were generated using identical definitions across sites. Sites executed the phenotyping workflow locally and produced deidentified aggregate outputs for pooled analysis while preserving local data governance.

Algorithm Validation Through Distributional Robustness

In the absence of site-specific EHR review, we assessed algorithm validity through distributional robustness testing.[28,29] Consistent performance across demographically divergent populations provides evidence of robustness,[30,31] whereas systematic miscalibration would be expected to manifest as divergent prevalence estimates across sites with different demographic compositions.[32]

The 58 hospitals across 4 regions represent substantially different populations in terms of racial and ethnic composition, comorbidity burden, and health care system scale, providing a rigorous test of algorithm generalizability beyond the development site.

Statistical Analysis

Continuous variables are reported as mean (SD) and compared using analysis of variance. Categorical variables are reported as counts and percentages and compared using χ^2 tests. Standardized residuals (SRes) from contingency tables identified organ systems driving distributional differences; values exceeding 1.96 indicate significance at $P < .05$, values exceeding 2.58 indicate significance at $P < .01$, and values exceeding 3.29 indicate significance at $P < .001$.

We calculated 95% CIs for prevalence using the exact binomial method. Cumulative PASC prevalence was calculated as the running total of PASC cases divided by the total COVID-19 cases through each quarter, expressed as a percentage. Temporal trends were modeled using negative binomial regression with the natural logarithm of quarterly COVID-19 cases as an offset, yielding incidence rate ratios (IRRs) with 95% CIs. Analyses were performed in R statistical software version 4.3.1 (R Project for Statistical Computing).

Results

COVID-19 Population Statistics

The study included a total of 457 950 patients with COVID-19 (mean age, 52.05 years; 275 107 [60.07%] female), including 153 880 in New England, 5002 in Southeast Texas, 46 419 in Southern California, and 252 649 in Western Pennsylvania (eTable 2 in Supplement 2). Patient age was similar across 3 of the regions (mean [SD], 57.54 [12.29] years in Southeast Texas, 57.71 [17.55] years in New England, and 58.01 [16.85] Southern California), while Western Pennsylvania had a notably lower mean (SD) age of 47.40 (24.27) years. The proportion of female patients was comparable across sites: 57.88% (26 867 patients) in Southern California, 59.10% (149 316 patients) in Western Pennsylvania, 62.22% (95 744 patients) in New England, and 63.57% (3180 patients) in Southeast Texas. Marked differences in race and ethnicity distributions were observed: New England (116 110 patients [75.45%]) and Western Pennsylvania (206 098 patients [81.57%]) had the highest proportion of White patients, Southeast Texas had the largest proportion of Black patients (1127 patients [22.53%]), and Southern California had the highest percentages of Asian (4905 patients [10.57%]) and Hispanic (7395 patients [15.93%]) patients.

PASC Prevalence

Among COVID-19 cases, the P2RC algorithm identified 74 560 PASC cases (16.28% overall: 28 585 [18.58%] in New England, 978 [19.55%] in Southeast Texas, 10 534 [22.69%] in Southern California, and 34 463 [13.64%] in Western Pennsylvania) (**Table 1**), which is consistent across 3 of the 4 regions, with Western Pennsylvania showing a notably lower prevalence. Compared with the overall COVID-19 population, patients with PASC were older, had higher comorbidity burden, and were more likely to be female. Race and ethnicity distributions among patients with PASC broadly mirrored those of the overall COVID-19 population at each site.

Organ System Distribution

Systemic symptoms were the most common manifestations across all sites (22.58%-25.09%), followed by respiratory (14.11%-19.09%) and gastrointestinal symptoms (13.28%-16.96%) (eFigure 2 in Supplement 2). Gynecologic and pelvic manifestations were least common (0.76%-1.38%).

The χ^2 analysis (**Figure 1**) revealed significant differences in organ system distribution across regions ($\chi^2 = 2504.73$; $P < .001$) (eFigures 3-14 in Supplement 2 for each organ). Endocrine manifestations provided the clearest window into this heterogeneity. Endocrine PASC was underrepresented in New England (SRes = −34.0; $P < .001$) and overrepresented in Southeast Texas

(SRes = 8.2; $P < .001$), Southern California (SRes = 17.8; $P < .001$), and Western Pennsylvania (SRes = 19.2; $P < .001$).

Compositional analysis revealed distinct phenotypic patterns (**Figure 2**). In New England, pancreatic manifestations (predominantly abnormal glucose and hyperglycemia, with minimal diabetes) represented 67.24% of endocrine cases and thyroid manifestations 32.76%, with dyslipidemia essentially absent. Southeast Texas and Southern California showed stronger metabolic predominance (pancreas: 87.25% and 81.02%; dyslipidemia: 1.47% and 8.01%; thyroid: 11.27% and 10.97%). Western Pennsylvania exhibited an intermediate pattern, with pancreatic manifestations representing 74.09% of endocrine cases, thyroid 20.19%, and dyslipidemia 5.71%. Notably, prediabetes was essentially absent in New England but was the leading pancreatic diagnosis in Southeast Texas (58.99% of pancreatic cases), Southern California (73.04%), and Western Pennsylvania (30.77%), whereas abnormal glucose predominated in New England (72.51% of pancreatic cases). Thyroid PASC, predominantly hypothyroidism, was markedly overrepresented in New England (SRes = 16.1; $P < .001$) (eFigure 8 in Supplement 2). Detailed phenotyping of other organ systems is presented in eFigures 15 to 25 in Supplement 2.

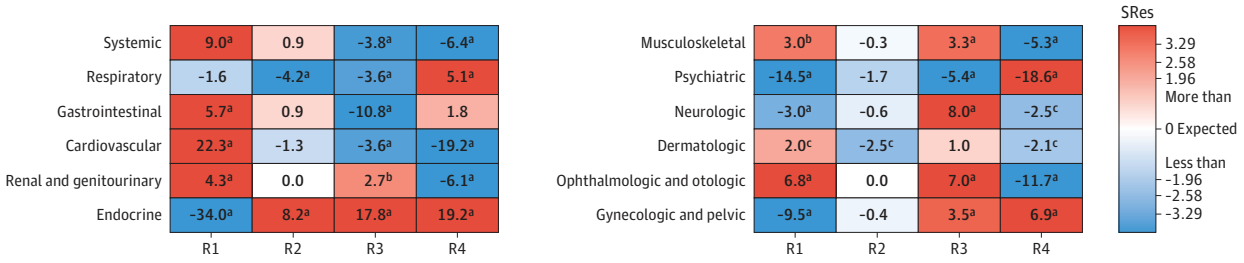
Table 1. Demographic and Clinical Characteristics of PASC Cases Identified by the Algorithm Across 4 Regions

Characteristic	Cases, No. (%)				*P* value
	New England	Southeast Texas	Southern California	Western Pennsylvania	
PASC incidence	47 827 (24.49)	1094 (5.35)	17 240 (30.93)	52 682 (18.87)	NA
Patients	28 585 (18.58)	978 (19.55)	10 534 (22.69)	34 463 (13.64)	NA
Age, mean (SD), y	61.06 (17.33)	59.48 (14.62)	61.86 (16.50)	55.12 (22.81)	<.001
Sex					
Female	18 257 (63.87)	645 (65.95)	6325 (60.04)	21 394 (62.08)	<.001
Male	10 328 (26.13)	333 (35.05)	4209 (29.96)	13 069 (37.92)	
Charlson Comorbidity Index score	3.10	2.82	3.45	3.26	NA
No. of unique Clinical Classifications Software Refined categories	36	38	53	52	NA
Race					
Asian	950 (3.32)	0	1036 (9.83)	370 (1.07)	<.001
Black	2710 (9.48)	123 (12.58)	3376 (32.05)	50 (0.15)	
Unknown	2570 (8.99)	235 (24.03)	548 (5.2)	2942 (8.54)	
White	21 593 (75.54)	499 (51.02)	3586 (34.04)	28 860 (83.74)	
Other[a]	650 (2.27)	37 (3.78)	1959 (18.6)	2220 (6.44)	
Hispanic ethnicity	1404 (4.91)	71 (7.26)	1708 (16.21)	411 (1.19)	<.001

Abbreviations: NA, not applicable; PASC, postacute sequelae of SARS-CoV-2 infection.

[a] Other race includes American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, Middle Eastern or North African, multiple races, and other unspecified groups.

Figure 1. Heat Map of χ^2 Analysis of Organ-Specific Postacute Sequelae of SARS-CoV-2 Infection Distribution Across 4 Regions (Rs)



Standardized residuals (SRes) reveal distinct patterns across New England (R1), Southeast Texas (R2), Southern California (R3), and Western Pennsylvania (R4). Red indicates overrepresentation, and blue indicates underrepresentation compared with expected frequencies.

[a] $P < .05$.

[b] $P < .01$

[c] $P < .001$.

Chronic Disease Burden

Among 883 *ICD-10-CM* codes associated with PASC manifestations, 594 (67.27%) represented chronic or potentially chronic conditions (356 [40.32%] definitively chronic and 238 [26.95%] potentially chronic) with only 36 codes (4.07%) representing acute, self-limited conditions (eTable 1 in Supplement 1). The prevalence of chronic PASC varied across regions (**Table 2**): 17.02% of all patients with COVID-19 in New England (26 185 of 153 880 patients), 15.43% in Southeast Texas (772 of 5002 patients), 19.88% in Southern California (9227 of 46 419 patients), and 12.03% in Western Pennsylvania (30 403 of 252 649 patients), for a total of 66 587 patients, or 14.54% of 457 950 patients. Collectively, this suggests that 66 587 of 74 560 patients with PASC (89.31%) have developed chronic conditions requiring ongoing clinical management.

Temporal Trends

Between quarter 2 of 2020 and quarter 2 of 2024, cumulative PASC prevalence showed slight increases across all regions (**Figure 3**A). The prevalence reached 18.57% (95% CI, 18.37%-18.76%) in New England, 19.54% (95% CI, 18.45%-20.67%) in Southeast Texas, 22.50% (95% CI, 22.12%-22.89%) in Southern California, and 13.59% (95% CI, 13.45%-13.72%) in Western Pennsylvania.

Negative binomial regression revealed significant quarterly increases in New England (IRR, 1.01; 95% CI, 1.00-1.01; $P < .001$; 0.6% relative increase per quarter) and Southern California (IRR, 1.00; 95% CI, 1.00-1.01; $P < .001$; 0.4% relative increase per quarter), and Western Pennsylvania (IRR, 1.02; 95% CI, 1.01-1.02; $P < .001$; 1.5% relative increase per quarter) (eTable 3 in Supplement 2). Southeast Texas showed a similar but nonsignificant trend (IRR, 1.00; 95% CI, 1.00-1.01; $P = .07$), likely owing to a smaller sample size.

Quarterly incidence rates remained stable through 2022 (9.52%-28.87%) but began diverging in late 2023, with Southeast Texas reaching 60.78% (95% CI, 46.11%-74.16%) by quarter 1 of 2024 (Figure 3B). These findings indicate an accumulating rather than resolving disease burden.

Figure 2. Bar Graph of Distribution of Endocrine Postacute Sequelae of SARS-CoV-2 Infection Phenotypes Across 4 Regions (Rs)

Table 2. Chronic PASC Burden Across 4 Regions

Region	COVID-19 cases, No.	PASC cases, No. (%)	Chronic PASC cases, No. (%)	Chronic PASC cases, % of all COVID-19 cases
New England	153 880	28 585 (18.58)	26 185 (91.60)	17.02
Southeast Texas	5002	978 (19.55)	772 (78.94)	15.43
Southern California	46 419	10 534 (22.69)	9227 (87.59)	19.88
Western Pennsylvania	252 649	34 463 (13.64)	30 403 (88.22)	12.03
Total	457 950	74 560 (16.28)	66 587 (89.31)	14.54

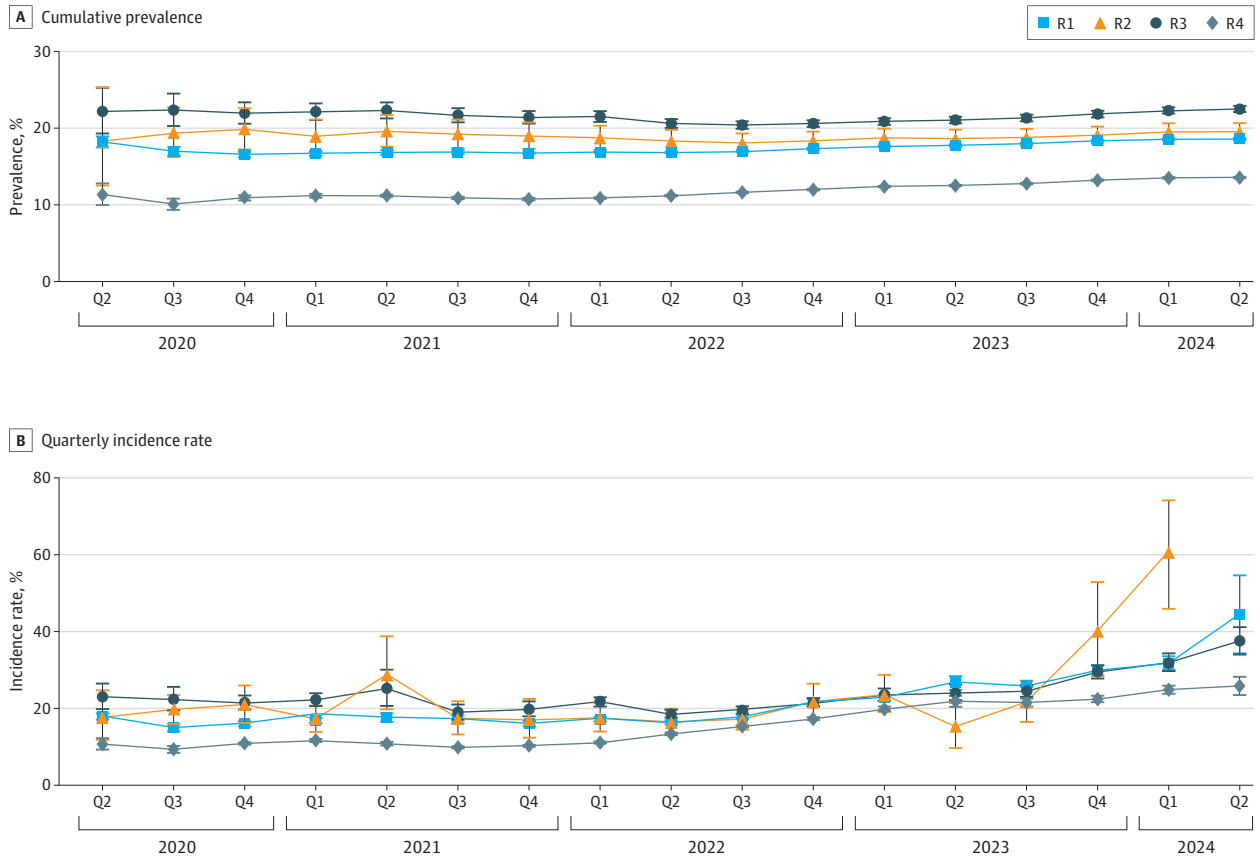
Abbreviation: PASC, postacute sequelae of SARS-CoV-2 infection.

Discussion

This multisite cohort investigation reveals systematic invisibility of a chronic disease epidemic following SARS-CoV-2 infection. Validated computable phenotyping identified PASC in 13.64% to 22.69% of patients with COVID-19 across 4 regions, consistent with recent population-based studies reporting long COVID prevalence of 10% to 26% among COVID-19 survivors using alternative ascertainment approaches.[33,34] Among identified *ICD-10-CM* codes, 67.27% involved chronic or potentially chronic conditions, with 14.54% of all patients with COVID-19 in 4 regions developing unexplained chronic conditions temporally associated with antecedent infection. We observed substantial heterogeneity in organ system involvement (χ^2 = 2504.73; P < .001), notably in endocrine manifestations, although the interpretation of these regional differences warrants caution. Temporal analyses demonstrated persistently increasing cumulative prevalence through mid-2024, revealing an accumulating burden rather than a resolving syndrome.

Prior studies demonstrate that U09.9 coding captures fewer than 1% of COVID-19 survivors,[16,35] while broader diagnostic code-based approaches identify approximately 7%.[16,17] Our phenotyping-derived estimates of 13.64% to 22.69%, thus represent a more than 2-fold improvement over the best-performing code-based surveillance, reflecting systematic underascertainment of chronic PASC within current public health infrastructure. Extrapolating our 14.54% chronic PASC prevalence to approximately 103 million documented US COVID-19 cases[36] suggests that 15 million individuals are living with chronic post–COVID-19 conditions, although this estimate should be interpreted

Figure 3. Line Graphs of Temporal Trends in Long COVID Across 4 Regions (Rs), 2020-2024



A, Cumulative prevalence of long COVID among all COVID-19 cases with 95% CIs (error bars). B, Quarterly incidence rate with 95% CIs (error bars), calculated as new long COVID cases divided by new COVID-19 cases in each quarter. Data are from New England (R1), Southeast Texas (R2), Southern California (R3), and Western Pennsylvania (R4). Quarters with fewer than 10 new COVID-19 cases were excluded from incidence rate calculations. Q indicates quarter.

cautiously, as our cohort requires longitudinal EHR documentation and excludes patients with preexisting conditions unlikely to represent PASC. This systematic underascertainment generates downstream consequences: health systems cannot accurately project capacity for ongoing care needs, disability programs cannot fully recognize the true burden, payers cannot develop appropriate reimbursement models, and investigators cannot efficiently recruit participants for therapeutic trials.

The temporal trajectory of PASC prevalence carries critical implications for health system planning. Our finding of persistently increasing cumulative prevalence through mid-2024 (4 years into the pandemic and well after widespread vaccination) contradicts assumptions that PASC represents a legacy of early, severe infection waves. The modest but statistically significant quarterly increases (IRR, 1.003-1.015) represent a 0.3% to 1.5% relative increase per quarter, compounded over a decade (40 quarters), and correspond to a 13% to 81% relative increase in cumulative PASC prevalence, underscoring the substantial long-term burden if current trends persist. This pattern reflects ongoing accrual of incident cases from successive infection waves rather than a fixed cohort progressing toward resolution. Because resolution of conditions cannot be reliably determined from retrospective EHR data, cumulative prevalence should be interpreted as the proportion of patients with COVID-19 who have ever been identified as having PASC, rather than as an estimate of current active cases. Health systems should anticipate sustained rather than diminishing demand for PASC-related care.

A natural question arising from these prevalence estimates is whether health systems should already be experiencing overwhelming demand from patients with PASC. We contend they are, but this burden manifests as unexplained increases in chronic disease management rather than as a discrete, labeled condition. Patients with chronic postviral conditions present to primary care with fatigue, to cardiology with dysautonomia, to endocrinology with new-onset diabetes, and to neurology with cognitive complaints, without the diagnostic code connecting these presentations to antecedent infection. Systematic underascertainment in surveillance systems does not mean these patients are absent from clinical care; rather, clinicians may recognize and manage post–COVID-19 conditions under alternative diagnostic codes, rendering the PASC burden invisible to population-level surveillance while remaining visible at the point of care. This fragmentation across specialty silos impedes both epidemiologic surveillance and coordinated clinical management, and may partly explain observed postpandemic increases in diabetes, cardiovascular disease, and fatigue syndromes.

That 67.27% of PASC manifestations represent chronic or potentially chronic conditions suggests PASC is better conceptualized as a chronic disease burden than as a self-limited postviral syndrome. Evidence suggests PASC patients generate approximately 1.5 times higher health care costs than matched controls,[37,38] with increased emergency department visits and specialist consultations. Applied to our prevalence estimates, this represents a substantial additional demand on already-strained systems.

The marked organ system heterogeneity we observed, thyroid-predominant in New England vs metabolic-predominant in Southeast Texas and Southern California, suggests distinct pathobiological mechanisms, although these patterns may also reflect differences in local coding practices or population-level comorbidity profiles. Should biological heterogeneity predominate, findings indicate PASC comprises multiple distinct postviral syndromes requiring precision phenotyping for precision medicine. Clinical trials should stratify by predominant organ system involvement rather than treating PASC as a monolithic entity,[39] and biomarker discovery efforts should identify molecular signatures distinguishing these endotypes,[40,41] enabling phenotype-specific interventions.

This investigation demonstrates that AI-powered federated phenotyping infrastructure is both feasible and scalable for multi-institutional PASC surveillance. The ENACT network's standardized i2b2 framework[42] with CCSR ontology mapping enabled reproducible deployment of this AI algorithm across 58 hospitals while preserving local data governance. Expanding this architecture

nationally would require additional sites to replicate the process demonstrated by our additional regions: implementing open-source R code within local i2b2 environments and executing standardized phenotyping workflows. However, refinements would enhance performance. The algorithm's attention mechanism may require site-specific fine-tuning, based on local EHR reviews, to optimize precision-recall trade-offs. Our study relied on validation through coherent prevalence estimates rather than a systematic EHR review, which limited our ability to detect site-specific miscalibration. Integration with existing public health surveillance systems would require the development of standardized data exchange specifications. Site-level *ICD-10-CM* code usage analysis within key CCSR categories, EHR review validation at additional sites, and matched comparator cohorts of uninfected individuals to quantify excess incidence are prioritized as next steps.

Limitations

This study has limitations that should be mentioned. Our phenotyping algorithm relies on the quality of EHR documentation, and the loyalty score filter selects for patients with sufficient longitudinal documentation; individuals with limited or fragmented health care engagement may, therefore, be underrepresented, and PASC prevalence in this population may, therefore, be underestimated. EHR review validation was not conducted in Southeast Texas, Southern California, and Western Pennsylvania, and site-specific precision and recall remain unquantified. However, the algorithm's consistent performance across 4 demographically distinct populations, ranging from a single academic medical center to a 40-hospital system, with a 50-fold variation in sample size, provides evidence of robustness under distributional shift, as systematic overcalling would be expected to produce greater prevalence variation.

The diagnosis-of-exclusion framework may miss cases among individuals with sparse health care engagement, and the 4 regions may not represent community practice patterns. Our clinical exclusion criteria may have inadvertently excluded true PASC cases in which preexisting conditions were exacerbated by COVID-19. Temporal association does not establish causation, and the absence of a COVID-19–negative comparator group precludes quantification of excess incidence above background rates; although the P2RC diagnosis-of-exclusion framework partially mitigates this concern, coincidental incident chronic disease cannot be fully excluded. The observed variation in dyslipidemia prevalence may reflect differences in local *ICD-10-CM* coding practices or population-level comorbidity patterns rather than true biological differences; site-level *ICD-10-CM* usage analysis and targeted EHR review would be required to disentangle these explanations.

We did not directly compare phenotyping-derived prevalence to site-specific U09.9 counts; instead, we relied on published estimates of U09.9 sensitivity (4.9%-19%) and population-level prevalence (<1%).[15-17,35] Site-specific comparisons would strengthen quantification of the surveillance gap but would not alter the central finding that validated phenotyping identifies substantially more cases than any code-based approach.

Conclusions

In this cohort study, precision phenotyping estimated that approximately 1 in 6 patients with COVID-19 developed PASC, and 89.31% of patients with PASC had at least 1 chronic condition requiring sustained clinical management. Current diagnostic coding captured only a fraction of affected individuals, leaving the majority invisible to surveillance systems. The substantial chronic PASC burden represented an accumulating health care crisis demanding urgent investment in surveillance infrastructure, integrated care pathways, and targeted therapeutics. Substantial organ system heterogeneity suggested the existence of distinct PASC endotypes, requiring precision medicine approaches with phenotype-specific interventions. The successful deployment of this custom AI algorithm across 58 hospitals in 4 regions demonstrates that federated AI infrastructure can transform PASC from invisible to code-based surveillance to actionable.

ARTICLE INFORMATION

Accepted for Publication: April 3, 2026.

Published: May 27, 2026. doi:10.1001/jamanetworkopen.2026.14909

Corresponding Author: Hossein Estiri, PhD, Department of Medicine, Massachusetts General Hospital, 399 Revolution Dr, Ste 790, Somerville, MA 02145 (hestiri@mgh.harvard.edu).

Author Affiliations: Department of Medicine, Massachusetts General Hospital, Boston (Tian, Azhir, Hügel, Cheng, Fard, Bassett, Klann, Estiri); Department of Medicine, Brigham and Women's Hospital, Boston, Massachusetts (Azhir); D. Bradley McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston (Decaro, Bernstam); CTSI/Biomedical Informatics Program, University of California, Los Angeles, Los Angeles (Chau); University Medical Center Göttingen, Department of Medical Informatics, Göttingen, Germany (Hügel); Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, Pennsylvania (Morris, Visweswaran); Department of Medicine, University of California, Los Angeles, Los Angeles (Bell); Department of Neurology, Massachusetts General Hospital, Boston (Murphy).

Author Contributions: Ms Tian and Dr Estiri had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Tian, Azhir, Hügel, Bassett, Bell, Murphy, Estiri.

Acquisition, analysis, or interpretation of data: Tian, Azhir, Decaro, Chau, Hügel, Morris, Cheng, Fard, Bell, Bernstam, Visweswaran, Klann, Estiri.

Drafting of the manuscript: Tian, Azhir, Fard, Estiri.

Critical review of the manuscript for important intellectual content: All authors.

Statistical analysis: Tian, Azhir, Decaro, Cheng, Fard, Estiri.

Obtained funding: Bernstam, Murphy, Estiri.

Administrative, technical, or material support: Tian, Chau, Hügel, Morris, Bassett, Bell, Bernstam, Visweswaran, Klann.

Supervision: Azhir, Bassett, Bernstam, Klann, Murphy, Estiri.

Conflict of Interest Disclosures: Dr Hügel reported receiving grants from the German Academic Exchange Service and the German Research Foundation during the conduct of the study. No other disclosures were reported.

Funding/Support: The National Institutes of Health supported the research reported in this publication under award number R01 AI165535 from the National Institute of Allergy and Infectious Diseases, and under award number U24 TR004111 from the National Center for Advancing Translational Sciences.

Role of the Funder/Sponsor: The funders 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; and decision to submit the manuscript for publication.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Sharing Statement: See Supplement 3.

REFERENCES

1. Sidky H, Young JC, Girvin AT, et al; N3C Consortium. Data quality considerations for evaluating COVID-19 treatments using real world data: learnings from the National COVID Cohort Collaborative (N3C). *BMC Med Res Methodol*. 2023;23(1):46. doi:10.1186/s12874-023-01839-2

2. Zhang HG, Honerlaw JP, Maripuri M, et al; Consortium for Clinical Characterization of COVID-19 by EHR (4CE). Potential pitfalls in the use of real-world data for studying long COVID. *Nat Med*. 2023;29(5):1040-1043. doi:10.1038/s41591-023-02274-y

3. Mandel HL, Shah SN, Bailey LC, et al; RECOVER EHR Cohort. Opportunities and challenges in using electronic health record systems to study postacute sequelae of SARS-CoV-2 infection: insights from the NIH RECOVER Initiative. *J Med Internet Res*. 2025;27:e59217. doi:10.2196/59217

4. Rudrapatna VA, Butte AJ. Opportunities and challenges in using real-world data for health care. *J Clin Invest*. 2020;130(2):565-574. doi:10.1172/JCI129197

5. Lin LY, Henderson AD, Carlile O, et al; OpenSAFELY Collaborative. Healthcare utilisation in people with long COVID: an OpenSAFELY cohort study. *BMC Med*. 2024;22(1):255. doi:10.1186/s12916-024-03477-x

6. Tarver ME. Race and ethnicity in real-world data sources: considerations for medical device regulatory efforts. *J Prim Care Community Health*. Published online March 6, 2021. doi:10.1177/21501327219940 40

7. Cook L, Espinoza J, Weiskopf NG, et al; N3C Consortium. Issues with variability in electronic health record data about race and ethnicity: descriptive analysis of the National COVID Cohort Collaborative Data Enclave. *JMIR Med Inform*. 2022;10(9):e39235. doi:10.2196/39235

8. Agampodi S, Tadesse BT, Sahastrabuddhe S, Excler JL, Kim JH. Biases in COVID-19 vaccine effectiveness studies using cohort design. *Front Med (Lausanne)*. 2024;11:1474045. doi:10.3389/fmed.2024.1474045

9. Millard LAC, Fernández-Sanlés A, Carter AR, et al. Exploring the impact of selection bias in observational studies of COVID-19: a simulation study. *Int J Epidemiol*. 2023;52(1):44-57. doi:10.1093/ije/dyac221

10. Ballering AV, van Zon SKR, Olde Hartman TC, Rosmalen JGM; Lifelines Corona Research Initiative. Persistence of somatic symptoms after COVID-19 in the Netherlands: an observational cohort study. *Lancet*. 2022;400 (10350):452-461. doi:10.1016/S0140-6736(22)01214-4

11. Bull-Otterson L, Baca S, Saydah S, et al. 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(21):713-717. doi:10.15585/mmwr.mm7121e1

12. Dagliati A, Strasser ZH, Hossein Abad ZS, et al; Consortium for Clinical Characterization of COVID-19 by EHR (4CE); Consortium for Clinical Characterization of COVID-19 by EHR (4CE). Characterization of long COVID temporal sub-phenotypes by distributed representation learning from electronic health record data: a cohort study. *EClinicalMedicine*. 2023;64:102210. doi:10.1016/j.eclinm.2023.102210

13. Chen C, Haupert SR, Zimmermann L, Shi X, Fritsche LG, Mukherjee B. Global prevalence of post-coronavirus disease 2019 (COVID-19) condition or long COVID: a meta-analysis and systematic review. *J Infect Dis*. 2022;226 (9):1593-1607. doi:10.1093/infdis/jiac136

14. Subramanian A, Nirantharakumar K, Hughes S, et al. Symptoms and risk factors for long COVID in non-hospitalized adults. *Nat Med*. 2022;28(8):1706-1714. doi:10.1038/s41591-022-01909-w

15. Maripuri M, Dey A, Honerlaw J, et al. Characterization of post-COVID-19 definitions and clinical coding practices: longitudinal study. *Online J Public Health Inform*. 2024;16:e53445. doi:10.2196/53445

16. Wang HI, Doran T, Crooks MG, et al. Prevalence, risk factors and characterisation of individuals with long COVID using electronic health records in over 1.5 million COVID cases in England. *J Infect*. 2024;89(4):106235. doi:10.1016/j.jinf.2024.106235

17. Fung KW, Baye F, Baik SH, Zheng Z, McDonald CJ. Prevalence and characteristics of long COVID in elderly patients: an observational cohort study of over 2 million adults in the US. *PLoS Med*. 2023;20(4):e1004194. doi:10.1371/journal.pmed.1004194

18. O'Mahoney LL, Routen A, Gillies C, et al. The prevalence and long-term health effects of long Covid among hospitalised and non-hospitalised populations: a systematic review and meta-analysis. *EClinicalMedicine*. 2022; 55:101762. doi:10.1016/j.eclinm.2022.101762

19. Fernandez-de-Las-Peñas C, Notarte KI, Macasaet R, et al. Persistence of post-COVID symptoms in the general population two years after SARS-CoV-2 infection: a systematic review and meta-analysis. *J Infect*. 2024;88 (2):77-88. doi:10.1016/j.jinf.2023.12.004

20. Azhir A, Hügel J, Tian J, et al. Precision phenotyping for curating research cohorts of patients with unexplained post-acute sequelae of COVID-19. *Med*. 2025;6(3):100532. doi:10.1016/j.medj.2024.10.009

21. Visweswaran S, Samayamuthu MJ, Morris M, et al. Development of a coronavirus disease 2019 (COVID-19) application ontology for the Accrual to Clinical Trials (ACT) network. *JAMIA Open*. 2021;4(2):ooab036. doi:10.1093/jamiaopen/ooab036

22. Morrato EH, Lennox LA, Dearing JW, et al. The Evolve to Next-Gen ACT Network: an evolving open-access, real-world data resource primed for real-world evidence research across the Clinical and Translational Science Award Consortium. *J Clin Transl Sci*. 2023;7(1):e224. doi:10.1017/cts.2023.617

23. Kohane IS, Churchill SE, Murphy SN. A translational engine at the national scale: informatics for integrating biology and the bedside. *J Am Med Inform Assoc*. 2012;19(2):181-185. doi:10.1136/amiajnl-2011-000492

24. Agency for Healthcare Research and Quality; Healthcare Cost and Utilization Project. Clinical Classifications Software Refined (CCSR). Accessed October 27, 2023. https://hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp

25. Klann JG, Henderson DW, Morris M, et al. A broadly applicable approach to enrich electronic-health-record cohorts by identifying patients with complete data: a multisite evaluation. *J Am Med Inform Assoc*. 2023;30(12): 1985-1994. doi:10.1093/jamia/ocad166

26. Hügel J, Sax U, Murphy SN, Estiri H. tSPM+; a high-performance algorithm for mining transitive sequential patterns from clinical data. *arXiv*. Preprint posted online September 8, 2023. doi:10.48550/ARXIV.2309.05671

27. Clai Group. Long COVID AI scripts. Accessed April 20, 2026. https://github.com/clai-group/long_covid_ai_scripts

28. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol*. 2016;69:245-247. doi:10.1016/j.jclinepi.2015.04.005

29. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol*. 2015;68(3):279-289. doi:10.1016/j.jclinepi.2014.06.018

30. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics*. 2020;21(2):345-352.

31. Bareinboim E, Pearl J. Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A*. 2016;113(27):7345-7352. doi:10.1073/pnas.1510507113

32. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. 2019;366(6464):447-453. doi:10.1126/science.aax2342

33. Mandel H, Yoo YJ, Allen AJ, et al. Long COVID incidence proportion in adults and children between 2020 and 2024: an electronic health record-based study from the RECOVER Initiative. *Clin Infect Dis*. 2025;80(6):1247-1261. doi:10.1093/cid/ciaf046

34. Shi J, Lu R, Tian Y, et al. Prevalence of and factors associated with long COVID among US adults: a nationwide survey. *BMC Public Health*. 2025;25(1):1758. doi:10.1186/s12889-025-22987-8

35. Lu Y, Lindaas A, Izurieta HS, et al. Lessons learned from characterizing long COVID among US Medicare beneficiaries. *Pharmacoepidemiol Drug Saf*. 2025;34(2):e70101. doi:10.1002/pds.70101

36. Centers for Disease Control and Prevention. Surveillance and data analytics: COVID. Accessed November 21, 2025. https://www.cdc.gov/covid/php/surveillance/index.html

37. Neba R, Pedaprolu LS, Neba B, Sambamoorthi U. Long COVID is associated with excess direct healthcare expenditures among adults in the United States. *Healthcare (Basel)*. 2025;13(21):2704. doi:10.3390/healthcare13212704

38. Mu Y, Dashtban A, Mizani MA, et al; CVD-COVID-UK/COVID-IMPACT Consortium. Healthcare utilisation of 282,080 individuals with long COVID over two years: a multiple matched control, longitudinal cohort analysis. *J R Soc Med*. 2024;117(11):369-381. doi:10.1177/01410768241288345

39. Thaweethai T, Jolley SE, Karlson EW, et al; RECOVER Consortium. Development of a definition of postacute sequelae of SARS-CoV-2 infection. *JAMA*. 2023;329(22):1934-1946. doi:10.1001/jama.2023.8823

40. Su Y, Yuan D, Chen DG, et al; ISB-Swedish COVID-19 Biobanking Unit. Multiple early factors anticipate post-acute COVID-19 sequelae. *Cell*. 2022;185(5):881-895.e20. doi:10.1016/j.cell.2022.01.014

41. Klein J, Wood J, Jaycox JR, et al. Distinguishing features of long COVID identified through immune profiling. *Nature*. 2023;623(7985):139-148. doi:10.1038/s41586-023-06651-y

42. Murphy SN, Weber G, Mendis M, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*. 2010;17(2):124-130. doi:10.1136/jamia.2009.000893

SUPPLEMENT 1.

eTable 1. *ICD-10* codes retained for PASC case identification with chronicity classifications.

SUPPLEMENT 2.

eFigure 1. Federated multi-site implementation workflow for PASC phenotyping across four regions.
eTable 2. Demographic and clinical characteristics of COVID-19 cases across four regions.
eFigure 2. Organ-specific PASC prevalence across four regions.
eFigure 3. Chi-square analysis of systemic PASC phenotype distribution across four regions.
eFigure 4. Chi-square analysis of respiratory PASC phenotype distribution across four regions.
eFigure 5. Chi-square analysis of gastrointestinal PASC phenotype distribution across four regions.
eFigure 6. Chi-square analysis of cardiovascular PASC phenotype distribution across four regions.
eFigure 7. Chi-square analysis of renal/genitourinary PASC phenotype distribution across four regions.
eFigure 8. Chi-square analysis of endocrine PASC phenotype distribution across four regions.
eFigure 9. Chi-square analysis of musculoskeletal PASC phenotype distribution across four regions.
eFigure 10. Chi-square analysis of psychiatric PASC phenotype distribution across four regions.
eFigure 11. Chi-square analysis of neurologic PASC phenotype distribution across four regions.
eFigure 12. Chi-square analysis of dermatologic PASC phenotype distribution across four regions.

SUPPLEMENT 3.
Data Sharing Statement