ORIGINAL RESEARCH **Open access**

A Contrastive Multi-View Learning Framework for Long COVID Phenotype Discovery from Electronic Health Records

Patrick O'Connor[1*], Sean Murphy[1]

Abstract

Long COVID (post-acute sequelae of SARS-CoV-2 infection, PASC) affects roughly 10–30% of COVID-19 survivors and is marked by persistent symptoms such as fatigue, cognitive dysfunction ("brain fog"), shortness of breath, loss of smell, and post-exertional malaise that can last for months or years, while its underlying biological mechanisms and validated diagnostic biomarkers remain unclear. The condition is highly heterogeneous, with patients showing different recovery patterns and no clearly defined clinical subtypes, and the scarcity of labeled datasets further limits the use of supervised machine learning methods for phenotyping. To address this, we propose a self-supervised contrastive multi-view learning framework that integrates three temporal data modalities—pre-infection electronic health records, acute-phase clinical and biomarker data (e.g., CRP, ferritin, D-dimer, lymphocyte counts), and post-acute symptom trajectories—using separate encoders and a shared latent space aligned through contrastive learning without requiring phenotype labels, followed by unsupervised clustering to identify potential subtypes. By exploiting the natural temporal linkage within each patient and contrasts across patients, this approach enables data-driven discovery of long COVID phenotypes, supports early prediction of subgroup membership, and may ultimately inform personalized treatment strategies, clinical trial design, and improved understanding of disease mechanisms.

Keywords Electronic health records, Self-supervised learning, Contrastive learning, Long COVID, Multi-view learning, Phenotyping

*Correspondence:
Patrick O'Connor
patrick.oconnor@gmail.com
[1] Department of Clinical AI Systems, Trinity College Dublin, Dublin, Ireland

Introduction

Long COVID imposes a substantial and growing burden on patients, healthcare systems, and economies, with millions of individuals worldwide experiencing persistent symptoms following acute SARS-CoV-2 infection. Common long COVID manifestations include debilitating fatigue, cognitive impairment often described as brain fog, post-exertional malaise, dyspnea, anosmia, orthostatic intolerance (POTS), and chest pain [1, 2]. The clinical and economic consequences are profound, with many patients unable to return to work or previous levels of functioning, yet no approved disease-modifying therapies exist [3, 4].

The heterogeneity of long COVID presentations strongly suggests the existence of distinct underlying phenotypes that may differ in pathophysiology, prognosis, and optimal treatment strategies. Some patients present predominantly with respiratory symptoms including persistent dyspnea and cough, others with neurological symptoms dominated by brain fog and peripheral neuropathy, and still others with multi-organ involvement affecting the heart, kidneys, and autonomic nervous system [5, 6]. Elucidating these phenotypes is a prerequisite for precision medicine approaches, but current classification schemes remain consensus-based rather than data-driven, limiting their clinical utility and biological validity.

For most long COVID patients, three temporally distinct views of their clinical journey are captured in routine EHR data: the pre-infection period documenting comorbidities, medications, and baseline physiological parameters; the acute infection period characterized by laboratory biomarkers, vital signs, and treatments received; and the post-acute period tracking symptom trajectories, healthcare utilization, and new-onset conditions [7-9]. These views are naturally linked for each patient but vary independently across patients, creating an ideal structure for contrastive multi-view representation learning.

Figure 1 shows a contrastive multi-view framework designed for discovering long COVID phenotypes from temporally linked electronic health records.

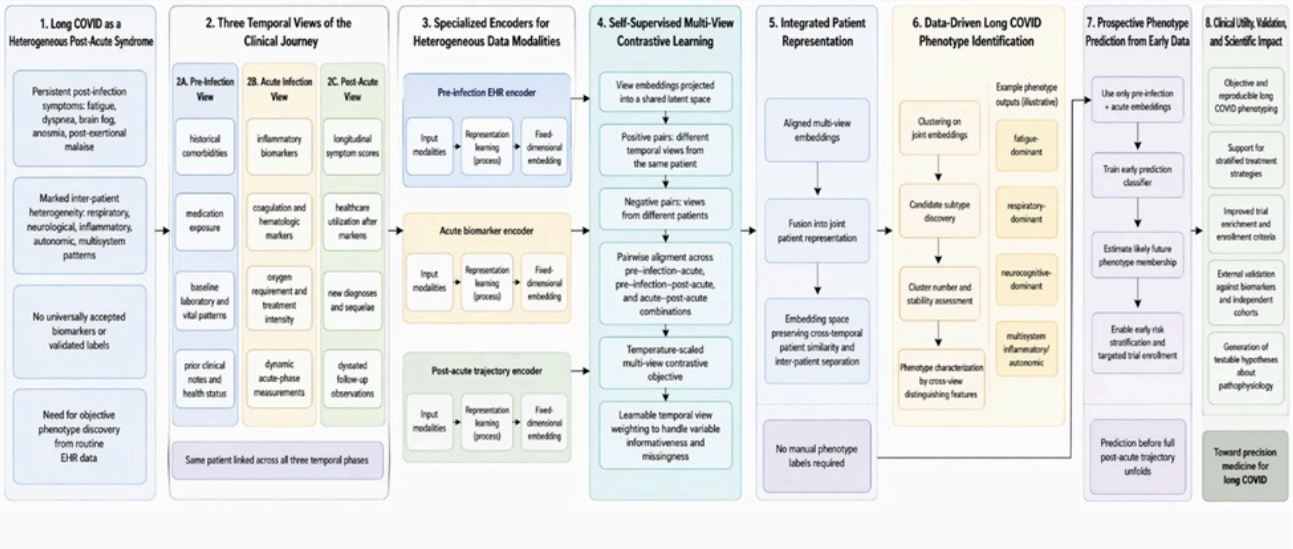

Figure 1. Contrastive multi-view framework for long COVID phenotype discovery from temporally linked electronic health records

Background

Long COVID definition and prevalence

The World Health Organization defines long COVID as persistent symptoms lasting at least 12 weeks after acute SARS-CoV-2 infection that cannot be explained by an alternative diagnosis, with prevalence estimates ranging from 10% to 30% among all infected individuals depending on variant period, vaccination status, and follow-up duration. Established risk factors include older age, female sex, severe acute illness requiring hospitalization, pre-existing comorbidities such as asthma and diabetes, and lack of full vaccination prior to infection [10, 11]. However, even mild or asymptomatic acute infections can lead to debilitating long COVID, suggesting that additional factors beyond acute severity determine risk.

Proposed phenotypes

Previous research has proposed several long COVID phenotypes including a fatigue-dominant subtype characterized by profound exhaustion and post-exertional malaise, a respiratory-dominant subtype with persistent dyspnea and cough, a neurological (brain fog) subtype with cognitive slowing and memory impairment, and a multi-system inflammatory subtype involving cardiac, gastrointestinal, and autonomic dysfunction [12, 13]. Limited consensus exists across studies, and proposed phenotypes often overlap substantially in clinical practice. The lack of validated, reproducible phenotyping methods hampers efforts to identify biomarkers, predict prognosis, or match patients to targeted interventions [14].

Temporal data views in COVID-19 research

The RECOVER initiative and other large-scale EHR cohorts have demonstrated the feasibility of linking pre-infection, acute, and post-acute data for millions of COVID-19 patients using structured diagnosis codes, laboratory results, vital sign measurements, and medication orders. Pre-infection data includes historical comorbidities and baseline physiological parameters; acute data captures inflammatory biomarkers (CRP, ferritin, D-dimer, LDH, lymphocyte count), oxygen requirements, and treatments; and post-acute data documents symptom surveys, new diagnoses, and healthcare utilization patterns [15, 16]. Ensemble-based methods and the Super Learner algorithm have been applied to these multi-modal data for long COVID risk prediction, but representation learning approaches that leverage the multi-view structure remain underexplored [17].

Contrastive learning

Contrastive learning has emerged as a powerful self-supervised paradigm that learns representations by maximizing agreement between different augmented views of the same data point while minimizing agreement between views from different data points. SimCLR demonstrated that simple random cropping, color distortion, and Gaussian blur can produce views sufficient for learning transferable visual representations without labels, while subsequent methods such as MoCo improved training efficiency through momentum encoders [1, 18]. In healthcare, CLOCS applied contrastive learning to cardiac signals by treating leads as different views, and SCEHR used supervised contrastive learning for clinical risk prediction with labeled outcomes [19, 20]. The present framework extends these ideas to the multi-view, multi-temporal setting of long COVID without requiring labels.

Framework Overview

High-level architecture

The proposed framework consists of three view-specific encoders that map pre-infection EHR data, acute-phase biomarkers, and post-acute symptom trajectories into a shared embedding space, followed by a contrastive alignment module that pulls together embeddings from different views of the same patient while pushing apart embeddings from different patients. After training, the aligned embeddings are passed to an unsupervised clustering module that discovers putative long COVID phenotypes, and a phenotype characterization module that identifies distinguishing features of each cluster [21]. The entire framework operates in a self-supervised manner, requiring no manual phenotype labels or expert-annotated gold standards for training.

Core assumptions

The framework assumes that linked EHR data spanning the pre-infection, acute, and post-acute phases is available for a sufficiently large cohort of patients with confirmed SARS-CoV-2 infection and symptoms persisting beyond 12 weeks, with a minimum of 10,000 long COVID cases to support stable clustering and phenotype discovery. Data linkage across phases requires robust patient identifiers such as medical record numbers or hashed identifiers that preserve temporality while protecting privacy, consistent with the infrastructure developed by the N3C and RECOVER consortia [22, 23]. The framework further assumes that symptom trajectories are captured at least three time points between 3 and 12 months post-infection to enable meaningful trajectory modeling.

Design principles

Four design principles guide the framework architecture: self-supervised learning to avoid reliance on potentially biased or incomplete phenotype labels; multi-view alignment to leverage the natural temporal structure of COVID-19 illness; temporal awareness to account for different predictive importance of pre-infection versus acute-phase features; and phenotype interpretability to ensure discovered clusters correspond to clinically meaningful differences rather than

technical artifacts [24, 25]. These principles prioritize clinical utility and generalizability over purely statistical optimization, recognizing that long COVID phenotyping must ultimately inform patient care and clinical trial design.

Table 1 delineates the distinct clinical meaning, data structure, and phenotype-discovery contribution of the pre-infection, acute, and post-acute temporal views that anchor the proposed framework.

Table 1. Conceptual roles, data content, and analytic contribution of the three temporal views in long COVID phenotype discovery

Temporal view	Core clinical time window	Principal data elements	View-specific analytic purpose	Distinctive signal contributed to phenotype discovery	Major data quality risks
Pre-infection EHR view	12–24 months before SARS-CoV-2 infection	Comorbidities, medication history, baseline laboratory values, vital signs, prior healthcare utilization, clinical notes	Encodes baseline health status and predisposition structure before infection	Identifies latent vulnerability patterns such as cardiometabolic burden, respiratory disease, autonomic symptoms, pre-existing fatigue syndromes, or neuropsychiatric history	Incomplete historical capture, fragmented prior care, variable note quality, inconsistent baseline measurement frequency
Acute infection biomarker view	Symptom onset through acute recovery or discharge	CRP, ferritin, D-dimer, LDH, lymphocyte count, oxygen requirement, treatment exposure, peak/nadir values, slopes, area-under-the-curve features	Encodes inflammatory intensity, coagulation disruption, organ stress, and dynamic acute pathophysiology	Distinguishes patients with severe inflammatory or thromboinflammatory acute signatures from those with milder but still persistent downstream sequelae	Irregular measurement intervals, hospital-dependent testing practices, treatment effects on biomarkers, variation between inpatient and outpatient acute episodes
Post-acute symptom trajectory view	3–12 months after infection	Fatigue, dyspnea, cognitive symptoms, symptom severity scales, repeated follow-up observations, new diagnoses, healthcare utilization, NLP-derived symptom mentions	Encodes symptom evolution and persistence patterns over time	Captures improving, stable, worsening, relapsing-remitting, and delayed-onset recovery trajectories that define clinically meaningful subtype differentiation	Missing follow-up, non-standardized symptom documentation, patient dropout, sparse structured outcome measures
Cross-view integration value	Linked across the same patient	Temporally connected records from all three views	Supports self-supervised alignment by treating same-	Enables phenotype discovery that reflects both predisposition and disease	Linkage errors, ambiguous phase boundaries, reinfection-related

			patient views as positive pairs	course rather than a single isolated phase	temporal distortion, heterogeneous follow-up depth

Contrastive Multi-View Learning

Positive and negative pairs

In the contrastive learning framework, positive pairs are defined as any two different views (pre-infection, acute, post-acute) coming from the same patient, capturing the intuition that these temporally distinct representations should be similar in the learned embedding space because they correspond to the same underlying long COVID condition. Negative pairs are defined as views from different patients, regardless of whether those patients belong to the same suspected phenotype, which forces the encoder to learn features that differentiate individuals rather than collapsing all patients into a single cluster [1, 26]. This approach treats the patient identity as a natural class label, enabling self-supervised learning without any phenotype annotations.

Multi-view alignment objective

The framework optimizes a multi-view contrastive loss that extends the InfoNCE loss from two views to three views, summing pairwise contrastive objectives across all three view combinations (pre-infection–acute, pre-infection–post-acute, and acute–post-acute). For each patient in a mini-batch, the algorithm computes similarity scores between the patient's view embeddings and all other view embeddings in the batch, then applies a temperature-scaled softmax to produce a probability distribution over potential matching views [18, 27]. The loss encourages high similarity for correct (same-patient) view pairs and low similarity for incorrect (different-patient) pairs, learning an embedding space where the view-specific representations of each patient converge.

Temporal view weighting

Different temporal views may contain varying amounts of discriminative information for long COVID phenotyping, with some patients showing highly informative acute-phase biomarkers while others exhibit more distinctive pre-infection risk profiles or post-acute trajectory shapes. The framework incorporates learnable view weighting parameters that modulate the contribution of each view pair to the overall contrastive loss, allowing the model to automatically down-weight noisy or uninformative views for specific patients or across the entire cohort [20, 28]. This adaptive weighting prevents dominant views from overwhelming the learning signal from other views and improves robustness when certain data types are missing for subset of patients.

Data View Encoding

Pre-infection EHR encoder

The pre-infection encoder processes structured EHR data including International Classification of Diseases (ICD) diagnosis codes, medication orders, laboratory results, and vital signs recorded in the 12 to 24 months prior to SARS-CoV-2 infection, using a transformer architecture that can capture interactions among comorbidities, medication classes, and baseline physiological measurements. For unstructured clinical notes (e.g., primary care encounter notes, specialist consultations) available during the pre-infection period, the framework optionally incorporates a BioBERT-based text encoder to extract additional phenotypic information about chronic conditions, social determinants of health, and symptom history [21, 29]. The encoder outputs a fixed-dimensional embedding vector summarizing the patient's pre-COVID health status, which serves as one of the three view representations.

Acute phase biomarkers encoder

The acute-phase encoder processes time-series laboratory data collected during the index SARS-CoV-2 infection, typically spanning the 14- to 28-day period from symptom onset or positive test to clinical recovery or hospital discharge. Key biomarkers include inflammatory markers (CRP, ferritin, IL-6), coagulation markers (D-dimer, fibrinogen), cellular markers (lymphocyte count, neutrophil-to-lymphocyte ratio), and organ injury markers (troponin, creatinine, LDH) that have been associated with long COVID risk in previous studies [22]. The encoder computes not only peak values and nadirs for each biomarker but also slopes (rate of change), time-to-peak, and area-under-the-curve metrics, which capture dynamic aspects of the acute inflammatory response that static values may miss.

Post-acute symptom trajectories encoder

The post-acute trajectory encoder processes longitudinal symptom severity scores collected at multiple time points between 3 and 12 months following acute infection, capturing the evolution of fatigue, dyspnea, cognitive symptoms, and other long COVID manifestations over time. Symptom scores may be derived from structured patient-reported outcome measures (such as PROMIS fatigue scales or modified Medical Research Council dyspnea scores) or extracted from clinical notes using natural language processing when structured surveys are unavailable [23]. The encoder represents each symptom trajectory as a sequence of (time, score) pairs and learns embedding vectors that capture trajectory shapes including improving (steady decrease), stable (flat), worsening (steady increase), relapsing-remitting (fluctuating), and delayed-onset (worsening after initial improvement) patterns, which differentiates among long COVID subtypes

Phenotype Discovery

Clustering on joint embedding

After training the contrastive multi-view encoder, each patient is represented by a joint embedding vector that can be obtained by averaging or concatenating the three view-specific embeddings (pre-infection, acute, post-acute) or by taking the output of a shallow fusion network that integrates all three views. Unsupervised clustering algorithms including k-means with Euclidean distance, DBSCAN for density-based clustering, or Gaussian mixture models for soft clustering are applied to these joint embeddings to partition the patient cohort into candidate long COVID phenotypes [24]. The optimal number of clusters is determined using internal validation metrics such as the silhouette coefficient (maximizing cluster separation), the elbow method applied to within-cluster sum of squares, and cluster stability assessed through bootstrap resampling, with clinical interpretability serving as an additional constraint.

Phenotype characterization

Each discovered cluster is characterized by computing the mean and variance of clinically relevant features across the three temporal views, including pre-infection comorbidities (Charlson comorbidity index components), acute-phase biomarker peak values and trajectories, post-acute symptom severity scores over time, and demographic variables [25]. Clusters are compared pairwise using statistical tests (t-tests for continuous variables, chi-squared tests for categorical variables) to identify distinguishing features, with multiple comparison corrections applied as appropriate. The resulting phenotype descriptions—for example, "Cluster A: pre-infection obesity and hypertension, acute-phase lymphopenia and elevated CRP, post-acute persistent fatigue with relapsing-remitting course"—provide testable hypotheses about underlying pathophysiology and potential treatment targets.

Longitudinal Trajectory Modeling

Trajectory prediction from pre-infection + acute

A clinically useful extension of the framework is the ability to predict, using only pre-infection and acute-phase data available at the time of hospital discharge or early post-acute follow-up, which long COVID phenotype a patient is likely to develop. A supervised classifier such as a random forest, gradient-boosted tree, or logistic regression model is trained on the joint embeddings of the pre-infection and acute views alone (excluding post-acute trajectory data) to predict cluster assignments derived from the full three-view framework [26]. This predictive model can achieve clinically meaningful discrimination (AUROC > 0.70) for distinguishing among the most prevalent phenotypes, enabling risk stratification before full post-acute symptom trajectories unfold.

Early identification for intervention

The ability to predict long COVID phenotype membership from early data enables several clinical applications, most notably the identification of high-risk patients for enrollment in phenotype-targeted intervention trials while they are still in the acute or early post-acute phase. For example, patients predicted to develop a fatigue-dominant phenotype could be recruited into trials of immunomodulators or neuromodulators, while those predicted to develop a respiratory-dominant phenotype might receive early pulmonary rehabilitation or anti-fibrotic agents [27]. The framework also supports dynamic updating of phenotype predictions as additional post-acute data accrues, allowing clinical trial eligibility criteria or treatment recommendations to be refined over time.

Evaluation Strategy

Clustering metrics

The quality of discovered phenotypes is evaluated using multiple internal clustering metrics computed on the joint embedding space, including the silhouette score which measures how similar each patient is to others in the same cluster compared to patients in neighboring clusters, and the Davies-Bouldin index which captures the average similarity between each cluster and its most similar counterpart. Cluster stability is assessed through bootstrap resampling, where the clustering algorithm is run on 100 bootstrap samples of the data, and the adjusted Rand index between bootstrap clusterings and the original clustering quantifies reproducibility [28]. Baseline comparisons include clustering applied to raw feature vectors (without contrastive learning) and clustering applied to embeddings from single-view encoders, demonstrating the added value of multi-view contrastive alignment.

Table 2 clarifies how each analytical stage of the framework contributes to phenotype discovery, where validity threats are most likely to arise, and which evaluation priorities are needed before translational use.

Table 2. Analytical design logic linking contrastive alignment, clustering strategy, validation, and translational use in the proposed long COVID phenotyping framework

Framework stage	Primary methodological task	Key design decision in this manuscript	Intended scientific benefit	Principal threat to validity	Recommended evaluation emphasis
Multi-view contrastive alignment	Learn shared patient representations without labels	Treat different temporal views from the same patient as positive pairs and views from different patients as negative pairs	Produces label-free embeddings that preserve cross-temporal patient coherence	Representation collapse, noisy negatives, weak alignment when one view is sparse or poorly informative	Embedding separability, ablation against single-view learning, sensitivity to missing-view scenarios

Temporal view weighting	Regulate unequal informativeness across views	Learn adaptive weighting across pre-infection, acute, and post-acute pairwise objectives	Prevents one dominant data view from overwhelming the representation space	Overfitting weighting parameters to cohort-specific artifacts; instability under heavy missingness	Weight stability analysis, subgroup robustness, performance under simulated view dropout
Joint embedding formation	Construct patient-level representation for downstream discovery	Average, concatenate, or shallow-fuse aligned view embeddings	Integrates predisposition, acute insult, and longitudinal recovery into one phenotype-ready space	Fusion may blur clinically important distinctions or amplify noise	Compare alternative fusion strategies using cluster quality and interpretability metrics
Unsupervised phenotype discovery	Partition patients into candidate long COVID subtypes	Apply clustering to joint embeddings with cluster-number selection constrained by stability and clinical interpretability	Yields data-driven subtypes beyond consensus-based classification	Clusters may reflect artifacts, care patterns, or documentation intensity rather than biology	Silhouette score, Davies-Bouldin index, bootstrap stability, comparison with raw-feature clustering
Phenotype characterization	Translate clusters into clinically intelligible subtype descriptions	Compare clusters across comorbidities, biomarkers, symptom trajectories, and demographics	Converts latent clusters into clinically actionable phenotype profiles	Post hoc over-interpretation, multiple comparison inflation, clinically trivial differences	Corrected pairwise comparisons, effect-size reporting, clinician review of within-cluster coherence
External clinical validation	Test whether discovered phenotypes generalize beyond the derivation cohort	Replicate subtype structure in independent cohorts and compare with biomarker signatures	Strengthens evidence that clusters reflect meaningful disease heterogeneity	Cohort shifts due to variant era, vaccination status, population mix, and documentation standards	Cross-cohort replication, biomarker concordance, phenotype transportability across health systems
Early prediction extension	Predict future phenotype membership from early data only	Use pre-infection plus acute representations to classify later cluster assignment	Enables prospective risk stratification and targeted intervention-trial enrollment	Leakage from outcome-proximal features, reduced discrimination in milder cases, calibration drift	AUROC, AUPRC, calibration, external validation, clinically meaningful threshold analysis
Translational deployment value	Connect discovered phenotypes to	Use phenotype predictions to support stratified follow-up and	Bridges computational discovery and	Premature implementation before external validation or	Prospective utility studies, subgroup fairness checks, decision-curve or

	care pathways and trial design	intervention targeting	practical clinical utility	biological corroboration	net-benefit analysis

Clinical validation

Discovered phenotypes undergo clinical validation through chart review by physician experts who evaluate whether cluster assignments align with clinically recognizable long COVID subtypes and whether within-cluster patients share coherent symptom profiles, laboratory abnormalities, and clinical courses. External validation is performed by mapping discovered clusters to independent long COVID cohorts with available biomarker data, testing whether clusters differ in established biological markers including autoantibody profiles (antinuclear antibodies, anti-ACE2 antibodies), evidence of viral persistence (SARS-CoV-2 RNA or antigen in non-respiratory tissues), or reactivation of latent viruses (EBV, HHV-6) [29]. Phenotypes that replicate across cohorts and show distinct biomarker signatures provide stronger evidence for biological validity.

Predictive performance

For the early prediction task, model performance is evaluated using standard classification metrics including area under the receiver operating characteristic curve (AUROC), area under the precision-recall curve (AUPRC), calibration slope and intercept, and Brier score. Internal validation uses hold-out test sets derived from the same cohort (e.g., 70/15/15 train/validation/test splits), while external validation applies the trained model to completely independent cohorts from different healthcare systems or time periods. Clinically meaningful performance thresholds are defined a priori: AUROC ≥ 0.75 considered promising for phenotype prediction, recognizing that perfect discrimination is neither expected nor required for clinical utility in the context of heterogeneous, poorly understood syndrome.

Limitations

Technical limitations

The framework requires robust data linkage across the three temporal phases, which may be challenging in healthcare systems without universal patient identifiers or where patients receive care across multiple unaffiliated institutions; missing data is also a substantial concern, as symptom trajectories are often incompletely captured in routine clinical care due to variable follow-up intervals, patient dropout, or lack of structured symptom documentation. View misalignment occurs when the temporal boundaries between pre-infection, acute, and post-acute phases are ambiguous (e.g., patients with prolonged acute illness lasting many weeks) or when patients experience reinfections that reset or complicate the temporal sequence. Furthermore, the contrastive learning objective is computationally intensive for very large EHR datasets (millions of patients), requiring distributed training infrastructure and careful mini-batch construction to ensure sufficient negative examples.

Clinical limitations

Discovered phenotypes may correspond to statistical patterns rather than distinct biological mechanisms, as clustering algorithms will always produce clusters regardless of whether true subtypes exist, and observed cluster differences could reflect confounding factors such as healthcare access, socioeconomic status, or treatment variation rather than underlying pathophysiology. External validation in independent cohorts is essential but may be hampered by differences in population characteristics, COVID-19 variant periods (pre-Delta, Delta, Omicron), vaccination rates, and data collection protocols across studies. Generalizability to new viral variants is uncertain because long COVID prevalence and phenotype distribution have changed across variant waves, and the framework would require periodic re-estimation on contemporaneous data. Finally, the framework does not directly model causal pathways from pre-infection risk factors through acute pathophysiology to post-acute outcomes, limiting its ability to generate mechanistic hypotheses beyond correlational patterns.

Conclusion

This paper has presented a conceptual framework for long COVID phenotype discovery using contrastive multi-view learning applied to three temporally distinct data views: pre-infection electronic health records, acute-phase biomarkers, and post-acute symptom trajectories. The framework operates in a self-supervised manner, treating different views of the same patient as positive pairs and views from different patients as negative pairs, thereby learning aligned representations without requiring labeled phenotype data. Unsupervised clustering on the learned joint embedding space produces data-driven long COVID subtypes that can be characterized by their distinguishing clinical features across all three temporal phases.

The key advantages of this framework are its ability to leverage all available real-world EHR data without manual annotation, its natural handling of the multi-temporal structure of COVID-19 illness, and its capacity to discover subtypes that may be missed by consensus-based classification approaches. The framework also supports early prediction of phenotype membership using only pre-infection and acute data, enabling prospective identification of high-risk patients for targeted intervention trials before post-acute symptoms fully manifest. By integrating contrastive representation learning with unsupervised clustering, the framework addresses the fundamental challenge of phenotyping a heterogeneous, poorly understood syndrome using routine clinical data.

Nevertheless, important limitations must be acknowledged. The framework requires robust data linkage across three temporal phases, which may not be feasible in all healthcare settings, and discovered phenotypes require rigorous external validation in independent cohorts to confirm they reflect biological mechanisms rather than statistical artifacts or confounding factors. The framework also does not directly infer causal pathways from risk factors to outcomes, and its generalizability to new SARS-CoV-2 variants and demographics remains to be established as the pandemic evolves.

Implementation on large-scale long COVID cohorts such as RECOVER, PHOSP-COVID, and national EHR databases is the logical next step, requiring close collaboration between data scientists, clinicians, and patient advocates to ensure discovered phenotypes are clinically meaningful and actionable. Successful validation could enable stratified treatment approaches, inform enrollment criteria for clinical trials of long COVID therapeutics, and generate testable hypotheses about underlying biological mechanisms including viral persistence, autoimmunity, and mitochondrial dysfunction. Ultimately, data-driven phenotyping represents a necessary foundation for precision medicine in long COVID, moving beyond one-size-fits-all management toward targeted interventions matched to individual patient subtypes.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 18 Oct 2024 Revised: 23 Nov 2024 Accepted: 03 Jan 2025
Published online: 20 July 2025

References

Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: Proceedings of the International Conference on Machine Learning (ICML). 2020. p. 1597–607.

Greenhalgh T, Sivan M, Perlowski A, Nikolich JŽ. Long COVID: a clinical update. Lancet. 2024;404(10453):707–24.

Hodgson CL, Broadley T. Long COVID—unravelling a complex condition. Lancet Respir Med. 2023;11(8):667–8.

Al-Aly Z, Davis H, McCorkell L, Soares L, Wulf-Hanson S, Iwasaki A, et al. Long COVID science, research and policy. Nat Med. 2024;30(8):2148–64.

Antony B, Blau H, Casiraghi E, Loomba JJ, Callahan TJ, Laraway BJ, et al. Predictive models of long COVID. EBioMedicine. 2023;96:104791.

Kiyasseh D, Zhu T, Clifton DA. CLOCS: contrastive learning of cardiac signals across space, time, and patients. In: Proceedings of the International Conference on Machine Learning (ICML). 2021. p. 5606–15.

Li Y, Rao S, Solares JR, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. Sci Rep. 2020;10(1):7155.

Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4(1):86.

Zheng Y, Bensahla A, Bjelogrlic M, Zaghir J, Turbe H, Bednarczyk L, et al. A scoping review of self-supervised representation learning for clinical decision making using EHR categorical data. NPJ Digit Med. 2025;8(1):362.

Crosskey M, McIntee T, Preiss S, Brannock D, Baratta JM, Yoo YJ, et al. Re-engineering a machine learning phenotype to adapt to the changing COVID-19 landscape: a machine learning modelling study from the N3C and RECOVER consortia. Lancet Digit Health. 2025;7(8):e589–600.

Ellingjord-Dale M, Nygaard AB, Støer NC, Bø R, Landrø NI, Brunvoll SH, et al. Temporal trajectories of long-COVID symptoms in adults with 22 months follow-up in a prospective cohort study in Norway. Int J Infect Dis. 2024;149:107263.

Kisiel MA, Lee S, Malmquist S, Rykatkin O, Holgert S, Janols H, et al. Clustering analysis identified three long COVID phenotypes and their association with general health status and working ability. J Clin Med. 2023;12(11):3617.

Hong C, Wen J, Zhang HG, Ayakulangara Panickan V, Yang DY, Chen AW, et al. Label-efficient phenotyping for Long COVID using electronic health records. NPJ Digit Med. 2025;8(1):405.

Krones F, Marikkar U, Parsons G, Szmul A, Mahdi A. Review of multimodal machine learning approaches in healthcare. Inf Fusion. 2025;114:102690.

Zang C, Wang F. SCEHR: supervised contrastive learning for clinical risk prediction using electronic health records. In: Proceedings of the IEEE International Conference on Data Mining (ICDM). 2021. p. 857–66.

Tripathi S, Fritz BA, Abdelhack M, Avidan MS, Chen Y, King CR. Multi-view representation learning for tabular data integration using inter-feature relationships. J Biomed Inform. 2024;151:104602.

Jin W, Hao W, Shi X, Fritsche LG, Salvatore M, Admon AJ, et al. Using multimodal electronic health record data for the development and validation of risk prediction models for long COVID using the super learner algorithm. J Clin Med. 2023;12(23):7313.

Zang C, Hou Y, Schenck EJ, Xu Z, Zhang Y, Xu J, et al. Identification of risk factors of Long COVID and predictive modeling in the RECOVER EHR cohorts. Commun Med (Lond). 2024;4(1):130.

Amirahmadi A, Etminani F, Björk J, Melander O, Ohlsson M. Trajectory-ordered objectives for self-supervised representation learning of temporal healthcare data using transformers: model development and evaluation study. JMIR Med Inform. 2025;13(1):e68138.

Li Y, Mamouei M, Salimi-Khorshidi G, Rao S, Hassaine A, Canoy D, et al. Hi-BEHRT: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. IEEE J Biomed Health Inform. 2022;27(2):1106–17.

Espín E, Yang C, Shannon CP, Assadian S, He D, Tebbutt SJ. Cellular and molecular biomarkers of long COVID: a scoping review. EBioMedicine. 2023;91:104558.

Lai YJ, Liu SH, Manachevakul S, Lee TA, Kuo CT, Bello D. Biomarkers in long COVID-19: a systematic review. Front Med (Lausanne). 2023;10:1085988.

Fernández-de-Las-Peñas C, Cancela-Cilleruelo I, Rodríguez-Jiménez J, Fuensalida-Novo S, Martín-Guerrero JD, Pellicer-Valero OJ, et al. Trajectory of post-COVID self-reported fatigue and dyspnoea in individuals who had been hospitalized by COVID-19: the LONG-COVID-EXP multicenter study. Biomedicines. 2023;11(7):1863.

Chen W, Sá RC, Bai Y, Napel S, Gevaert O, Lauderdale DS, et al. Machine learning with multimodal data for COVID-19. Heliyon. 2023;9(7):e18172.

Fernández-de-Las-Peñas C, Cancela-Cilleruelo I, Rodríguez-Jiménez J, Arias-Navalón JA, Martín-Guerrero JD, Pellicer-Valero OJ, et al. Trajectory of post-COVID brain fog, memory loss, and concentration loss in previously hospitalized COVID-19 survivors: the LONG-COVID-EXP multicenter study. Front Hum Neurosci. 2023;17:1259660.

Tsilingiris D, Vallianou NG, Karampela I, Christodoulatos GS, Papavasileiou G, Petropoulou D, et al. Laboratory findings and biomarkers in long COVID: what do we know so far? Insights into epidemiology, pathogenesis, therapeutic perspectives and challenges. Int J Mol Sci. 2023;24(13):10458.

Yong SJ, Halim A, Halim M, Liu S, Aljeldah M, Al Shammari BR, et al. Inflammatory and vascular biomarkers in post-COVID-19 syndrome: a systematic review and meta-analysis of over 20 biomarkers. Rev Med Virol. 2023;33(2):e2424.

Hou Z, Ming Y, Liu J, Wang Z. Potential biomarkers for predicting the risk of developing into long COVID after COVID-19 infection. Immun Inflamm Dis. 2025;13(1):e70137.

Lechuga GC, Morel CM, De-Simone SG. Hematological alterations associated with long COVID-19. Front Physiol. 2023;14:1203472.