

Long COVID undercoding and early prediction of its documentation: survey comparison and an exportable model in the All of Us Research Program

Amir Asiaee¹
Kaveh Aryan²

¹Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

²Department of Informatics, King's College London, London, UK

Correspondence: amir.asiaeetaheri@vumc.org

Abstract

Objective. Long COVID often goes uncoded (U09.9). We estimated coding by group and built an exportable model predicting later documentation from early electronic health record (EHR) data.

Materials and Methods. In the All of Us Research Program, we followed 23,790 adults with COVID-19 for 455 days and compared coded shares with Household Pulse Survey prevalence. Among 23,300 adults uncoded at day 90, models predicted a first U09.9 code in days 91 to 455 (260 coded), evaluated by cross-validation and at held-out sites (EHR source groups). These included naive Bayes and linear discriminant models fit from released summary counts, and we implemented the naive Bayes model as a SMART on FHIR application.

Results. About 1 in 9 adults who ever had Long COVID carried the code, fewer among women than men (coding ratio 0.79, 95% sensitivity interval 0.65–0.95). At day 90, an elastic net predicted a subsequent code with AUROC 0.68 (0.64–0.71). At pooled held-out sites, AUROCs across models were 0.62–0.67. Adjusted for utilization, high scores predicted new Long COVID–associated symptom codes (rate ratios 1.7–2.1) no more than fractures and lacerations (1.9–2.0). On a validation split, the best summary-count model, a shrunken linear discriminant, matched the participant-level elastic net (AUROC 0.68 for both), although the elastic net had higher AUPRC. At held-out sites, a separately fitted 41-feature summary-count naive Bayes model reached AUROC 0.66 (0.58–0.72), against 0.62 for the elastic net (difference 0.04, 95% CI –0.03 to 0.11).

Discussion. Survey comparison quantifies undercoding in populations. Individual disease probabilities need further assumptions and validation.

Conclusion. Under stated assumptions, Long COVID is substantially and unevenly under-coded. Early EHR data modestly predict its documentation, and models fit from released summary counts, which can leave a data enclave, discriminate about as well as participant-level models.

Keywords: Long COVID; positive-unlabeled learning; electronic health records; underdiagnosis; clinical decision support

1 Background and Significance

Long COVID is defined clinically by symptoms that persist or emerge after SARS-CoV-2 infection. In the electronic health record (EHR), clinicians can record it with the ICD-10-CM code U09.9, introduced on October 1, 2021, but the presence of the code is not a clinical reference standard. Clinicians use the code far less often than the condition occurs. Its sensitivity was 0.15 against the highly symptomatic RECOVER research-index phenotype [1] and 10.5% to 14.9% against chart review in two health systems [2]. Its use also differs by clinician, site, sex, race, and insurance [3–8]. Case finding, registries, and trial screening that rely on the code therefore miss patients whose Long COVID is not coded.

Machine-learning models for Long COVID have mostly used the code itself or a clinic visit as the outcome [9–12]. In the All of Us Research Program, a supervised model reached an area under the receiver operating characteristic curve (AUROC) of 0.736 [13], and the National COVID Cohort Collaborative (N3C) model reached 0.72 when run without retraining [14]. Such models describe who gets diagnosed. Because they treat every uncoded patient as a negative example, they learn the coding habits of a health system.

Positive-unlabeled (PU) learning treats coded records as labeled positives and all other records as unlabeled, and reading those labels as confirmed disease requires an additional assumption [15, 16]. The EHR alone cannot identify its central quantity, the probability that a true case is labeled [17]. Federal surveys, however, measure Long COVID prevalence directly, by demographic group [18]. In this paper, we compare the two sources to estimate this coding fraction, under assumptions about population comparability and label validity that we state and vary. A second obstacle is practical. Research enclaves such as All of Us do not allow models trained on participant-

level rows to leave the enclave, so a model must be rebuilt from summary statistics that meet dissemination rules before clinicians can use it [19]. Methods exist for fitting models from shared summaries or across sites without pooling rows [20–24], but few reach a deployed EHR application, and a gap between published and running models persists even then [25]. We therefore also fit a model from released summary counts and implement it in an EHR application.

2 Objective

In this paper, we have three aims. First, we estimate the fraction of Long COVID that receives a diagnosis code, by sex, age, and race and ethnicity, from federal survey prevalence. Second, we build and validate models that predict a subsequent code at fixed times after infection, and we check whether their scores track Long COVID-associated documentation. Third, we export a model trained only on summary counts and implement it as a standards-based EHR application.

3 Materials and Methods

3.1 Data sources

We used the All of Us Research Program Registered Tier Dataset v9 (curated data repository R2025Q4R6, data cutoff January 1, 2025), accessed through the Researcher Workbench [26], in which dates are shifted backward by 1 to 365 days per participant, preserving intervals within a person. Survey prevalence came from the CDC/NCHS Household Pulse Survey (HPS, data.cdc.gov dataset gsea-w83j): 27 periods from June 2022 to September 2024, with the share of adults who had COVID-19 who ever had, or currently have, Long COVID [18]. Prices came from 2024 CMS Medicare physician payment data.

3.2 Cohort and outcomes

We defined the index date as the first ICD-10-CM U07.1 code on or after shifted date July 2, 2021. Because true dates are never earlier than shifted dates, day 91 after every index falls on or after the date U09.9 became billable. We required age 18 years or older, no Long COVID code on or before the index, and an EHR observation period from 365 days before to 455 days after the index

(Figure 1). Requiring follow-up through day 455 selected patients with sustained EHR coverage, using information unavailable at the prediction landmark. We took sex at birth from the All of Us participant record.

We define two outcomes. For the coding-fraction analysis, S is any U09.9 code in days 1 to 455 in the full cohort. For prediction at landmark day L , we exclude patients with a Long COVID code (U09.9 or the SNOMED CT post-COVID concepts) by day L (eligibility E_L), and S_L is a first U09.9 code in days $L + 1$ to 455. Our primary landmark is day 90. We grouped participants by the source identifier associated with their index event, which we call the site (54 sources, Supplementary Table S9). Of the day-90 cohort, 6.6% had a participant-mediated EHR source.

3.3 Features

Features used only information available by the landmark, in windows relative to the index (days -365 to -91 , -90 to -1 , 0 to 30 , and 31 to L). They covered demographics, acute severity, visit counts and their change from baseline, new symptom codes in 16 clusters and their number, a repeat COVID-19 code, prior conditions, tests after infection, and new medications (102 features at day 90, Supplementary Table S1). Features use source vocabularies (ICD-10-CM, CPT, LOINC, RxNorm). In every analysis, we took ICD-10-CM codes from both the condition and the observation tables (the data model files some codes, such as R53.83, as observations). The primary model used whether a laboratory test was recorded, not its result, because result units differ across sources and reference ranges are often missing (secondary analysis in Supplementary Section S6).

3.4 Coding fraction

Let Y denote Long COVID as defined by the survey, S the code, and g a demographic group, and define the coding fraction $c_g = P(S = 1 \mid Y = 1, g)$. If codes are assigned only to people with Long COVID and the survey prevalence π_g equals the prevalence in the corresponding All of Us group, then $P(S = 1 \mid g) = c_g \pi_g$, so

$$\hat{c}_g = \hat{p}_g / \hat{\pi}_g, \tag{1}$$

where $\hat{p}_g$ is the share of group g coded in days 1 to 455. This population identity does not require coding to be random within a group. More generally, with $q_g = P(Y = 1 \mid S = 1, g)$, the positive

predictive value of the code, and $\kappa_g = \pi_g^{\text{AoU}} / \pi_g^{\text{HPS}}$, the ratio of cohort to survey prevalence,

$$c_g = \frac{q_g p_g}{\kappa_g \pi_g^{\text{HPS}}}. \quad (2)$$

The primary estimate sets $q_g = \kappa_g = 1$. We report $\hat{c}_g$ and the coding ratio $\hat{c}_g / \hat{c}_{\text{ref}}$ by sex, age band, and race and ethnicity, under the “ever” and “current” survey definitions, computing coded shares from counts rounded up to a multiple of 10. We report central 95% sensitivity intervals from 5,000 replicates, each drawing one survey period for all groups and a binomial draw of each $\hat{p}_g$. These intervals combine variation across survey periods with binomial variation and are not confidence intervals for the mean-period estimate. Sensitivity analyses varied q and κ , survey error, composition reweighting, and age weights (Supplementary Section S2).

3.5 Prediction models

At landmark L , the target is $f_L(x) = P(S_L = 1 \mid X_L = x, E_L)$, the probability of a subsequent code. We did not observe disease directly. We fit five models: prevalence only (M0), logistic regression on utilization features (M1), elastic-net logistic regression (M2, with ℓ_1 ratio 0.5 and $C = 0.05$), gradient boosting (M3), and bagged PU learning (M4), in which each of 25 bags contrasts the coded patients with a bootstrap sample of uncoded patients [16]. We evaluated logistic-model outputs as coded-outcome probabilities and the bagged PU model as a ranking score, because we did not recalibrate its sampled class distribution to cohort prevalence. We estimated performance by 5-fold cross-validation and on a test set of whole held-out sites (about a quarter of patients, Supplementary Table S8). We report the AUROC and the area under the precision-recall curve (AUPRC) for the coded outcome, with 95% bootstrap CIs (200 replicates, or 500 for the paired held-out-site comparison), and recall and positive predictive value among the top 5% of scores, computed from counts rounded up to a multiple of 10. Bootstrap intervals resampled participants with fixed predictions, so they do not incorporate model refitting, model selection, or variation across new sources.

The coded-outcome AUROC is the primary result. Supplementary Section S5 shows how it would relate to discrimination of disease under hypothetical disease prevalence in the day-90 population [27, 28].

Score adjustment. If coding among uncoded people with Long COVID does not depend on x within group g , then $P(Y = 1 \mid X_L = x, E_L) = f_L(x)/c_{g,L}$. Because the survey estimates the full-cohort c_g , not $c_{g,L}$, in an exploratory analysis we divided women’s scores by the women-to-men coding ratio (Supplementary Table S3).

Documentation checks. Among patients without a subsequent U09.9 code, we compared codes in days 91 to 455 between the top 10% and the bottom 50% of scores for 5 Long COVID–associated symptom groups and 4 negative-control conditions, crude and, for new codes, with Mantel–Haenszel adjustment across 12 utilization strata. *Lead time* was the number of days from the landmark to the first code among later-coded patients in the top 5% of scores.

Landmark tasks. We rebuilt the cohort, features, and outcome at $L \in \{30, 60, 90, 120, 180\}$ days and fit an elastic net at each [29]. These are different prediction tasks, and we restricted landmarks before day 90 to index dates for which every outcome day falls on or after October 1, 2021.

Subgroups and clinical utility. At day 90, we report discrimination, calibration-in-the-large, and the share in the top 5% of scores by sex, race and ethnicity, and age band (from a separate five-fold split). We also report decision curves [30] and, at a 5% review workload, the share of later-coded patients that the utilization model, a symptom-count rule, and the elastic net each found. Supplementary Section S3 analyzes related post-infectious conditions. Supplementary analyses compared alternative model families (Supplementary Section S7) and, exploratorily, the export-compatible binary features with continuous age and event counts (Supplementary Section S8). Reporting follows TRIPOD+AI [31], and Supplementary Table S4 lists the cohort, features, folds, and model used in each analysis.

3.6 A model that can leave the enclave

All of Us permits exporting a model only if it was trained on summary statistics that comply with its dissemination policy. On a random 80% training split we exported, for each binary feature, its 2×2 table with the outcome and, for each pair of kept features, the co-occurrence count. We kept a feature only if all four cells of its outcome table exceeded 20 and rounded every released count up to a multiple of 10 (we would have suppressed a pair count of 1 to 20, but none occurred). Outside the enclave we subtracted the expected rounding excess (4.5) from each count and fit eight candidates, naive Bayes and linear discriminants with and without shrinkage (Supplementary Section S6). We

imported the weights and evaluated them on the remaining 20% (the validation split) alongside participant-level models refit on the same training split. Both model families can leave the enclave, because both are computed from the same released counts. For the application, we implemented naive Bayes with empirical-Bayes shrinkage, because each of its weights is a per-feature log odds ratio that the application can show as that finding’s contribution to the score. We calibrated it by logistic regression on binned training-split scores (each bin with at least 50 coded and 50 uncoded patients).

To evaluate the procedure at sites that contributed no data to fitting, we repeated it with the 18 held-out sites excluded from fitting. Counts from 70% of training-site patients gave a shrunken naive Bayes score, exported score bins from the other 30% (at least 21 coded and 21 uncoded per bin) gave its calibration map, and we evaluated both at the held-out sites against an elastic net refit on all training-site patients. This evaluates a separately fitted 41-feature version. The implemented file contains 57 weighted features and 12 auxiliary feature definitions, and its calibration map has not undergone that evaluation. In a heuristic perturbation analysis, we refit the model 200 times with rounded counts replaced by compatible values.

3.7 Implementation

The exported model is a versioned JSON file (version 1.2.0) of feature definitions, weights, calibration map, and risk bands. A browser-only SMART on FHIR application [32] computes the features from the patient’s chart and shows the risk band and the dated findings behind the score, and a reference scorer and synthetic test patients define the expected output. The implemented model is the day-90 naive Bayes model, without score adjustment or landmark updating. The application applies the training cohort’s index and eligibility rules: the index is the first U07.1 condition on or after July 2, 2021, the patient must be 18 years or older at the index, and the chart may hold no Long COVID code by the scoring date, including before the index. The application scores charts from day 90 to day 455 after the index with the day-90 estimate.

3.8 Utilization before documentation

For patients later coded, we counted visits and tests from day 91 to the day before the first code and compared rates per 30 days with patients in the same stratum of age band, sex, acute

hospitalization, and baseline utilization who were not coded. We valued the difference at 2024 Medicare allowed amounts.

3.9 Disclosure control and ethics

We suppressed counts of 1 to 20, rounded other counts up to a multiple of 10, and, within each table, suppressed a further count wherever a reader could otherwise recover a suppressed count from the table’s totals (complementary suppression). All of Us participants provide informed consent at enrollment under a protocol overseen by the program’s institutional review board, and the program’s data passport model allows registered researchers to analyze de-identified Registered Tier data without project-specific review [\[verify\]](#).

4 Results

4.1 Cohort

Of 50,210 participants with a U07.1 code, 23,790 met all criteria (Table 1 and Supplementary Figure S1), and 23,300 had no Long COVID code by day 90. Of these, 260 (1.1%) received a first U09.9 code in days 91 to 455. Held-out sites had more women (69.4% vs 60.2%) and non-Hispanic White patients (73.6% vs 57.4%) than training sites (Supplementary Table S8).

4.2 Coding fractions

An estimated 1 in 9 adults who had ever had Long COVID, and 1 in 4 who currently had it, carried a U09.9 code (Table 2 and Figure 2). Women had higher survey prevalence than men (34.3% vs 23.9% ever) but a similar coded share (3.2% vs 2.9%), giving a coding ratio of 0.79 (95% interval 0.65–0.95) under the ever definition and 0.72 (0.58–0.88) under the current definition. The lower coding fraction for women persisted when we added within-period survey error (0.65–0.96, ever) and after we reweighted the survey to the cohort’s composition within each sex (0.76 ever, 0.75 current). The point ratio reached 1 at a relative PPV-to-prevalence multiplier of about 1.27 (ever) and 1.39 (current), and the upper sensitivity-interval limit at about 1.05 and 1.13 (Supplementary Section S2). Hispanic adults had a coding ratio of 0.79 (0.63–0.98, or 0.62–1.00 with within-period error) relative to non-Hispanic White adults under the ever definition only. Coding fractions did not

differ clearly by age band. Among the 18,830 adults whose index dates placed all of days 1 to 455 after U09.9 was introduced, the coded share was lower than in the full cohort (2.8%, Supplementary Section S2). Among the 10 sites with enough patients, the ratio of observed to expected coding ranged from 0.79 at the 10th percentile to 2.37 at the 90th.

4.3 Prediction of a subsequent code at day 90

The elastic-net model reached a cross-validated AUROC of 0.680 (95% CI 0.643–0.712) and an AUPRC of 0.034, three times the 1.1% prevalence. Gradient boosting and bagged PU learning reached 0.681 and 0.663, and the utilization-only model 0.574 (Table 3). On the 18 held-out sites pooled (5,760 patients, 60 coded), AUROCs ranged from 0.62 for the elastic net and 0.63 for the utilization-only model to 0.67 for bagged PU learning, with wide CIs. The elastic net did not outperform the utilization-only model there. Patients in the top 5% of elastic-net scores were coded at 4.1% (recall 19%). In cross-validation, alternative model families did not exceed the elastic net (Supplementary Section S7). Among later-coded patients in the top 5%, the score preceded the code by a median of 55 days (interquartile range 15–109). When we added abnormal laboratory results and vital signs, discrimination did not improve (cross-validated AUROC 0.682 vs 0.688 and held-out-site AUROC 0.624 vs 0.618, Supplementary Section S6), and richer features changed the cross-validated AUROC by 0.025 or less (Supplementary Section S8).

Among patients without a subsequent U09.9 code, those in the top decile had any later code for fatigue, dyspnea, cognitive, autonomic, and dizziness symptoms at 2.9 to 4.0 times the rate of the bottom half, and for fractures, lacerations, and cataract at 1.5 to 2.7 times (Figure 3B). For new codes, adjusted for utilization, the rate ratios were 1.7 to 2.1 for the 5 symptom groups and 1.9, 2.0, and 1.1 for fractures, lacerations, and cataract (Figure 3B). Appendicitis had too few events for a stable estimate.

4.4 Landmark tasks

Across landmark-specific tasks, AUROC ranged from 0.63 at day 30 to 0.70 at day 180 (Figure 4). In the same patients (eligible at day 180), the day-180 model’s scores had higher AUROC than the day-30 model’s (0.68 vs 0.62, a difference of 0.055, 95% CI 0.0001 to 0.113).

4.5 Subgroups and score adjustment

Coded-outcome AUROC estimates were 0.70 for women and 0.67 for men, and 0.68 to 0.70 across the 3 largest racial and ethnic groups, with overlapping intervals. These estimates do not establish equivalent performance, and they varied more by age (Supplementary Table S2). Dividing women’s scores by the women-to-men coding ratio raised the share of women among the top 5% from 67% to 75% (62% of the cohort). Recall was 19% before and 23% after the adjustment, one step of the rounding to multiples of 10 in the coded count, and AUROC changed little (0.688 and 0.684, Supplementary Table S3). At a 5% review workload, the symptom-count rule found 15% of later-coded patients, the utilization model 12%, and the elastic net 19%. The elastic net flagged 5.4% of women and 4.4% to 7.0% of the racial, ethnic, and age groups shown (Supplementary Table S2). The decision curve favored the model over flagging no one at all thresholds (0.5% to 5%) and over flagging everyone from 1% to 5% (Supplementary Figure S2).

4.6 The exported model

Of 102 features, 57 met the cell-size rule for export, and we released all 1,596 pair counts among them. On the validation split (4,680 patients, 50 coded), the best summary-count model was a shrunken linear discriminant ($\lambda = 0.6$, with empirical-Bayes shrinkage), with an AUROC of 0.68 (95% CI 0.60–0.76) and AUPRC of 0.035, against 0.68 (95% CI 0.59–0.76) and 0.089 for the elastic net refit on participant rows (Table 4). Naive Bayes with empirical-Bayes shrinkage, the model we implemented, had an AUROC of 0.67 (95% CI 0.59–0.75) and AUPRC of 0.028. The eight summary-count candidates were within 0.013 of one another in AUROC. The elastic net had higher AUPRC than any of them (0.028 to 0.051), with a wide interval (0.026–0.172) from only 50 coded patients. We did not test equivalence. Refit weights were stable under perturbation of the rounded counts (Supplementary Section S6). The largest weights were for a personal-history-of-COVID-19 code, a COVID-19 code at a follow-up visit in days 31 to 90, and a chest CT (Figure 5). When we excluded the 18 held-out sites from fitting, the separately fitted 41-feature naive Bayes score (counts from 12,310 training-site patients, 150 coded) had an AUROC of 0.66 (95% CI 0.58–0.72) at the held-out sites, against 0.62 (95% CI 0.52–0.70) for the elastic net refit on all training-site patients (a paired difference of 0.040, 95% CI -0.028 to 0.109). AUPRC was 0.018 against 0.025 (a difference

of -0.007 , 95% CI -0.027 to 0.001). Its calibration map, exported from two training-site bins, had a calibration slope of 1.02 , an intercept of -0.27 , and an observed-to-expected ratio of 0.90 at the held-out sites. We did not establish the precision of these calibration estimates. The implemented file contains 57 weighted features and a calibration map from 0.3% to 6.2% predicted probability of a subsequent code, against a post-COVID average of 1.0% . Its risk bands are multiples of that average and are descriptive score categories, not validated action thresholds.

4.7 Utilization before documentation

Between day 91 and the code (median 101 days), patients later coded had 6.6 (95% CI 3.7 – 9.5) more outpatient visits and 0.4 (0.2 – 0.7) more emergency visits than matched patients who were not coded, along with more echocardiograms, chest CT scans, and pulmonary function tests, a difference of about \$900 per patient at 2024 Medicare prices.

5 Discussion

Under the stated assumptions, about 1 in 9 adults who had ever had Long COVID, and 1 in 4 who currently had it, carried the diagnosis code, and the coding fraction was lower for women than men although women reported the condition more often. Chart data available 90 days after infection modestly predicted a subsequent code (AUROC 0.68), but the score tracked later documentation and utilization in general, not Long COVID–associated symptoms specifically. Models that we rebuilt from privacy-compliant summary counts discriminated about as well as the participant-level elastic net, both on a random validation split (AUROC 0.68 for the best summary-count model and for the elastic net) and at held-out sites (0.66 vs 0.62 , an imprecise difference), although the elastic net had higher AUPRC on the validation split. The same procedure produced the exported model that we implemented in an EHR application.

5.1 Comparison with prior work

Our estimated coding fractions are consistent with two direct measurements of U09.9 sensitivity: 0.15 against the highly symptomatic RECOVER research-index phenotype [1] and 10.5% to 14.9% against chart review [2]. They need no chart review but rest on the assumptions in Equation (2).

The lower coding fraction for women adds to reports of differences in Long COVID diagnosis by sex, race and ethnicity, insurance, and level of care [3, 6–8]. Here, we measure the gap against reported prevalence rather than inferring it from who is coded. In a companion analysis that linked self-reported diagnoses to participants’ records, we found lower crude coded shares for women than men for 5 common conditions, with ratios near 1 within EHR sites [33].

Published Long COVID models in All of Us reached AUROCs of 0.72 and 0.736 [13, 14]. Their outcome definitions and feature windows differ from our fixed day-90 prediction of a future code. N3C models with AUROCs above 0.85 predicted a Long COVID clinic visit [9]. The closest early-prediction comparator used history and the first 21 days after infection, extended to discharge for longer hospitalizations, and reported cross-site AUROCs of 0.74 to 0.75 overall and 0.60 to 0.66 among hospitalized patients [10]. Label-efficient phenotyping with chart-reviewed outcomes [2] addresses disease status directly. Our work complements it with a fixed prediction time, held-out sites, a survey-based estimate of undercoding, and a route from an enclave to an EHR application.

5.2 Implications

Diagnosis codes alone do not identify disease prevalence or the coding fraction without additional information or assumptions. Survey comparison supplies that information under stated assumptions, and the same construction could apply to other conditions that federal surveys measure. Exporting the model as summary counts addresses a barrier that keeps enclave research from reaching care. To run the model, another site needs the versioned JSON file, the reference scorer, and the synthetic test patients that define its expected output. No participant data and no model trained on participant rows leave the enclave. Compared with federated or distributed fitting [20, 23, 24], this route needs no network between sites but restricts the model to binary features with sufficient counts. Each weight is a readable log odds ratio, and the application shows the dated chart entries behind each flag. We intend the score for surfacing charts for clinician review or for registry and trial outreach. It is research decision support, not a diagnosis. Patients later coded used more care before their code than comparable uncoded patients. Whether earlier recognition changes utilization or spending requires prospective evaluation.

5.3 Limitations

First, the outcome is a recorded code, not a validated diagnosis. U09.9 has a positive predictive value of 38% to 64% against chart review or the RECOVER index [1, 2], so the no-false-positive assumption in the primary coding-fraction estimate is consequential. A code anywhere in 455 days can also refer to resolved disease, which makes the current-prevalence comparison an assumption-dependent benchmark. Second, survey prevalence comes from a low-response online survey of self-reported Long COVID, and the coding fraction depends on how the cohort differs from the survey population. For two well-coded conditions, the women-to-men ratio of cohort-to-national prevalence was 0.89 to 0.96, above the 0.79 that would remove the lower coding fraction for women (Supplementary Section S2). Third, discrimination at held-out sites was imprecise, the elastic net did not outperform a utilization-only model there, and we validated across source groups within one research network rather than in an external EHR. We defined sites by the index-event source, and 12.3% of held-out-site patients also had records from a training-site source by day 90. All of Us is not representative of the US population [34]. Fourth, we required follow-up through day 455 for inclusion, so performance may differ in patients with incomplete follow-up or care outside the contributing source. Fifth, the summary-count model excluded rare but specific findings (20 or fewer patients in a cell), and laboratory results were sparse and inconsistently coded, so the primary model used test records. Finally, we have tested the application only on synthetic patients: we have not validated its feature mapping from the research data model to FHIR resources on clinical records, participant-provided FHIR data in All of Us hold far fewer condition and procedure records per person than the research data [35], and the application reads FHIR administrative gender rather than sex at birth. A silent prospective evaluation would be the next step [25].

6 Conclusion

Under the stated assumptions, survey comparisons suggest substantial undercoding of Long COVID and a lower coding fraction among women. Early EHR features modestly predict later documentation. Models that we fitted from released summary counts discriminated about as well as participant-level models, and they can leave the enclave. Clinical use requires validation of the implemented model and its FHIR inputs.

Author contributions

Amir Asiaee: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft. Kaveh Aryan: Software, Writing – review and editing. [\[verify\]](#)

Competing interests

None declared. [\[verify\]](#)

Funding

This work received no specific funding. [\[verify\]](#)

Supplementary material

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

Use of AI tools

(Move to Acknowledgments for submission.) We used AI assistants, Claude (Anthropic) and Codex (OpenAI), for code development, analysis, and manuscript drafting and editing. Codex also carried out a pre-submission review of the manuscript and code in read-only sessions separate from development, and we reviewed and adjudicated each finding. All analyses of participant data ran inside the Researcher Workbench. Codex submitted scripts from the project repository to the Workbench through a restricted terminal interface that accepted only commands to install a checksum-verified script, run it, and return its aggregate outputs after automated disclosure checks. Neither assistant could view or export participant-level data. We reviewed all code, results, and text and take responsibility for the content. [\[verify\]](#)

Acknowledgments

We gratefully acknowledge All of Us participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health’s All of Us Research Program for making available the participant data examined in this study. The All of Us Research Program is supported by the National Institutes of Health, Office of the Director: Regional Medical Centers: 1 OT2 OD026549; 1 OT2 OD026554; 1 OT2 OD026557; 1 OT2 OD026556; 1 OT2 OD026550; 1 OT2 OD 026552; 1 OT2 OD026553; 1 OT2 OD026548; 1 OT2 OD026551; 1 OT2 OD026555; IAA #: AOD 16037; Federally Qualified Health Centers: HHSN 263201600085U; Data and Research Center: 5 U2C OD023196; Biobank: 1 U24 OD023121; The Participant Center: U24 OD023176; Participant Technology Systems Center: 1 U24 OD023163; Communications and Engagement: 3 OT2 OD023205; 3 OT2 OD023206; and Community Partners: 1 OT2 OD025277; 3 OT2 OD025315; 1 OT2 OD025337; 1 OT2 OD025276. In addition, the All of Us Research Program would not be possible without the partnership of its participants.

Data and code availability

This study used data from the All of Us Research Program’s Registered Tier Dataset v9, available to authorized users on the Researcher Workbench. Survey and Medicare data are public. The exported model, its specification, and synthetic test patients with expected outputs are available on the developer pages of <https://undiagnosednomore.org>. We will release the analysis code and the reference scorer in a public repository (URL and license at posting).

References

- [1] V. M. Castro, V. Gainer, N. Wattanasin, et al. Characterization and validation of EHR computable phenotypes for Long COVID using patient-reported symptoms: insights from the nationwide RECOVER program. *Journal of the American Medical Informatics Association*, 33(9):1759–1769, 2026. doi: 10.1093/jamia/ocag115.
- [2] C. Hong, J. Wen, H. G. Zhang, et al. Label efficient phenotyping for Long COVID using electronic health records. *npj Digital Medicine*, 8:405, 2025. doi: 10.1038/s41746-025-01617-y.

- [3] E. R. Pfaff, C. Madlock-Brown, J. M. Baratta, et al. Coding long COVID: characterizing a new disease through an ICD-10 lens. *BMC Medicine*, 21:58, 2023. doi: 10.1186/s12916-023-02737-6.
- [4] Nathaniel Hendrix, Rishi V. Parikh, Madeline Taskier, Grace Walter, Ilia Rochlin, Sharon Saydah, Emilia H. Koumans, Oscar Rincón-Guevara, David H. Rehkopf, and Robert L. Phillips. Heterogeneity of diagnosis and documentation of post-COVID conditions in primary care: A machine learning analysis. *PLOS One*, 20(5):e0324017, 2025. doi: 10.1371/journal.pone.0324017.
- [5] Monika Maripuri, Andrew Dey, Jacqueline Honerlaw, Chuan Hong, Yuk-Lam Ho, Vidisha Tanukonda, Alicia W Chen, Vidul Ayakulangara Panickan, Xuan Wang, Harrison G Zhang, Doris Yang, Malarkodi Jebathilagam Samayamuthu, Michele Morris, Shyam Visweswaran, Brendin Beaulieu-Jones, Rachel Ramoni, Sumitra Muralidhar, J Michael Gaziano, Katherine Liao, Zongqi Xia, Gabriel A Brat, Tianxi Cai, and Kelly Cho. Characterization of post-COVID-19 definitions and clinical coding practices: Longitudinal study. *Online Journal of Public Health Informatics*, 16:e53445, 2024. doi: 10.2196/53445.
- [6] Zachary Butzin-Dozier, Yunwen Ji, Sarang Deshpande, A. Jerrod Anzalone, Lin-Chiun Wang, Eric Hurwitz, Manav Kumar, Rena C. Patel, Jing Sun, McKaylee Robertson, Andrew Mertens, Marlene Camacho-Rivera, John M. Colford, Alan E. Hubbard, and Charisse Madlock-Brown. Black patients underdiagnosed with Long COVID in the US compared with White patients. *Health Affairs*, 45(7):775–781, 2026. doi: 10.1377/hlthaff.2025.00179.
- [7] Kevin Petranu, Tiffany Ford, Denise Henley, Ebenezer Hammond, John Henley, Melvin Crawford, Zella Coleman, and Jana Hirschtick. Exploring Long COVID diagnostic equity by sex, race and ethnicity, and insurance type. *Preventive Medicine Reports*, 65:103459, 2026. doi: 10.1016/j.pmedr.2026.103459.
- [8] Anne Rivelli, Jeffrey Ording, Julie Sheehan, and Jana L Hirschtick. Demographic differences in Long COVID diagnosis across levels of care: Implications for clinical practice. *Journal of General Internal Medicine*, 2026. doi: 10.1007/s11606-026-10389-4.
- [9] E. R. Pfaff, A. T. Girvin, T. D. Bennett, et al. Identifying who has long COVID in the USA: a

- machine learning approach using N3C data. *The Lancet Digital Health*, 4(7):e532–e541, 2022. doi: 10.1016/S2589-7500(22)00048-6.
- [10] Blessy Antony, Hannah Blau, Elena Casiraghi, Johanna J. Loomba, Tiffany J. Callahan, Bryan J. Laraway, Kenneth J. Wilkins, Corneliu C. Antonescu, Giorgio Valentini, Andrew E. Williams, Peter N. Robinson, Justin T. Reese, T.M. Murali, and N3C consortium. Predictive models of long COVID. *eBioMedicine*, 96:104777, 2023. doi: 10.1016/j.ebiom.2023.104777.
- [11] Chengxi Zang, Yu Hou, Edward J. Schenck, Zhenxing Xu, Yongkang Zhang, Jie Xu, Jiang Bian, Dmitry Morozuk, Dhruv Khullar, Anna S. Nordvig, Elizabeth A. Shenkman, Russell L. Rothman, Jason P. Block, Kristin Lyman, Yiye Zhang, Jay Varma, Mark G. Weiner, Thomas W. Carton, Fei Wang, and Rainu Kaushal. Identification of risk factors of Long COVID and predictive modeling in the RECOVER EHR cohorts. *Communications Medicine*, 4(1):130, 2024. doi: 10.1038/s43856-024-00549-0.
- [12] Miles Crosskey, Tomas McIntee, Sandy Preiss, Daniel Brannock, John M Baratta, Yun Jae Yoo, Emily Hadley, Frank Blanceró, Robert Chew, Johanna Loomba, Abhishek Bhatia, Christopher G Chute, Melissa Haendel, Richard Moffitt, and Emily R Pfaff. 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. *The Lancet Digital Health*, 7(8):100887, 2025. doi: 10.1016/j.landig.2025.100887.
- [13] C. Guardo, X. Zhang, S. Gangireddy, et al. Multi-scale data improves performance of machine learning model for long COVID identification. *Communications Medicine*, 6:389, 2026. doi: 10.1038/s43856-026-01621-7.
- [14] Emily R Pfaff, Andrew T Girvin, Miles Crosskey, Srushti Gangireddy, Hiral Master, Wei-Qi Wei, V Eric Kerchberger, Mark Weiner, Paul A Harris, Melissa Basford, Chris Lunt, Christopher G Chute, Richard A Moffitt, Melissa Haendel, and N3C and RECOVER Consortia. De-black-boxing health AI: demonstrating reproducible machine learning computable phenotypes using the N3C-RECOVER Long COVID model in the *All of Us* data repository. *Journal of the American Medical Informatics Association*, 30(7):1305–1312, 2023. doi: 10.1093/jamia/ocad077.

- [15] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 213–220, 2008. doi: 10.1145/1401890.1401920.
- [16] J. Bekker and J. Davis. Learning from positive and unlabeled data: a survey. *Machine Learning*, 109(4):719–760, 2020. doi: 10.1007/s10994-020-05877-5.
- [17] G. Ward, T. Hastie, S. Barry, J. Elith, and J. R. Leathwick. Presence-only data and the EM algorithm. *Biometrics*, 65(2):554–563, 2009. doi: 10.1111/j.1541-0420.2008.01116.x.
- [18] N. D. Ford, D. Slaughter, D. Edwards, et al. Long COVID and significant activity limitation among adults, by age — United States, June 1–13, 2022, to June 7–19, 2023. *MMWR Morbidity and Mortality Weekly Report*, 72(32):866–870, 2023. doi: 10.15585/mmwr.mm7232a3.
- [19] All of Us Research Program. Frequently asked questions. <https://support.researchallofus.org/hc/en-us/articles/52861503885460-Frequently-Asked-Questions>, 2026. Accessed October 8, 2026.
- [20] Yuan Wu, Xiaoqian Jiang, Jihoon Kim, and Lucila Ohno-Machado. Grid Binary LOGistic REGression (GLORE): building shared models without sharing data. *Journal of the American Medical Informatics Association*, 19(5):758–764, 2012. doi: 10.1136/amiajnl-2012-000862.
- [21] Amadou Gaye, Yannick Marcon, Julia Isaeva, Philippe LaFlamme, Andrew Turner, Elinor M Jones, Joel Minion, Andrew W Boyd, Christopher J Newby, Marja-Liisa Nuotio, Rebecca Wilson, Oliver Butters, Barnaby Murtagh, Ipek Demir, Dany Doiron, Lisette Giepmans, Susan E Wallace, Isabelle Budin-Ljøsne, Carsten Oliver Schmidt, Paolo Boffetta, Mathieu Boniol, Maria Bota, Kim W Carter, Nick deKlerk, Chris Dibben, Richard W Francis, Tero Hiekkalinna, Kristian Hveem, Kirsti Kvaløy, Sean Millar, Ivan J Perry, Annette Peters, Catherine M Phillips, Frank Popham, Gillian Raab, Eva Reischl, Nuala Sheehan, Melanie Waldenberger, Markus Perola, Edwin van den Heuvel, John Macleod, Bartha M Knoppers, Ronald P Stolk, Isabel Fortier, Jennifer R Harris, Bruce HR Woffenbuttel, Madeleine J Murtagh, Vincent Ferretti, and Paul R Burton. DataSHIELD: taking the analysis to the data, not the data to the analysis. *International Journal of Epidemiology*, 43(6):1929–1944, 2014. doi: 10.1093/ije/dyu188.

- [22] Nilanjan Chatterjee, Yi-Hau Chen, Paige Maas, and Raymond J. Carroll. Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association*, 111(513):107–117, 2016. doi: 10.1080/01621459.2015.1123157.
- [23] Ritoban Kundu, Maxwell Salvatore, Kumar Kshitij Patel, Lucila Ohno-Machado, Hyunghoon Cho, Xu Shi, and Bhramar Mukherjee. Privacy-enhancing sequential learning under heterogeneous selection bias in multi-site electronic health records data. *Journal of the American Medical Informatics Association*, 33(9):1679–1692, 2026. doi: 10.1093/jamia/ocag083.
- [24] Kelly Mohe Li, Jenna Marie Reps, Akihiko Nishimura, Martijn J. Schuemie, and Marc A. Suchard. Transfer-learning on federated observational healthcare data for prediction models using Bayesian sparse logistic regression with informed priors. *Journal of the American Medical Informatics Association*, 33(2):409–423, 2026. doi: 10.1093/jamia/ocaf146.
- [25] Conor K. Corbin, Rob Maclay, Aakash Acharya, Sreedevi Mony, Soumya Punnathanam, Rahul Thapa, Nikesh Kotecha, Nigam H. Shah, and Jonathan H. Chen. DEPLOYR: a technical framework for deploying custom real-time machine learning models into the electronic medical record. *Journal of the American Medical Informatics Association*, 30(9):1532–1542, 2023. doi: 10.1093/jamia/ocad114.
- [26] The All of Us Research Program Investigators. The “All of Us” Research Program. *New England Journal of Medicine*, 381(7):668–676, 2019. doi: 10.1056/NEJMs1809937.
- [27] S. Jain, M. White, and P. Radivojac. Recovering true classifier performance in positive-unlabeled learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1), 2017. doi: 10.1609/aaai.v31i1.10937.
- [28] R. Ramola, S. Jain, and P. Radivojac. Estimating classification accuracy in positive-unlabeled learning: characterization and correction strategies. In *Pacific Symposium on Biocomputing 2019*, pages 124–135, 2019. doi: 10.1142/9789813279827_0012.
- [29] Hans C. van Houwelingen. Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1):70–85, 2007. doi: 10.1111/j.1467-9469.2006.00529.x.

- [30] Andrew J. Vickers and Elena B. Elkin. Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006. doi: 10.1177/0272989x06295361.
- [31] Gary S Collins, Karel G M Moons, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten van Smeden, Anne-Laure Boulesteix, Jennifer Catherine Camaradou, Leo Anthony Celi, Spiros Denaxas, Alastair K Denniston, Ben Glocker, Robert M Golub, Hugh Harvey, Georg Heinze, Michael M Hoffman, André Pascal Kengne, Emily Lam, Naomi Lee, Elizabeth W Loder, Lena Maier-Hein, Bilal A Mateen, Melissa D McCradden, Lauren Oakden-Rayner, Johan Ordish, Richard Parnell, Sherri Rose, Karandeep Singh, Laure Wynants, and Patricia Logullo. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024. doi: 10.1136/bmj-2023-078378.
- [32] Joshua C Mandel, David A Kreda, Kenneth D Mandl, Isaac S Kohane, and Rachel B Ramoni. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. *Journal of the American Medical Informatics Association*, 23(5):899–908, 2016. doi: 10.1093/jamia/ocv189.
- [33] Amir Asiaee and Kaveh Aryan. Electronic Health Record Coding of Self-Reported Diagnoses in the All of Us Research Program, 2026. Companion manuscript; medRxiv preprint forthcoming.
- [34] Barrett Wallace Montgomery, Mahmoud Elkasabi, Michael Daniel Brannock, Laura Marcial, Ariba Huda, Melissa McPheeters, Claire Schulkey, Sarra Hedden, Philip Greenland, Chandan Sastry, Jennifer Adjemian, and Tamara R Litwin. Assessing the generalizability of prevalence estimates from the *All of Us* Research Program. *American Journal of Epidemiology*, 195(7):1962–1973, 2026. doi: 10.1093/aje/kwag010.
- [35] Jason Patterson, Elise Minto, Maura Beaton, Adit Anand, Mark Velez, Paul Harris, George Hripcsak, and Karthik Natarajan. A comparison of Fast Healthcare Interoperability Resources and Observational Medical Outcomes Partnership electronic health record data within the All of Us Research Program. *Journal of the American Medical Informatics Association*, 33(9):1593–1604, 2026. doi: 10.1093/jamia/ocag095.

Table 1: Cohort characteristics. Adults with EHR-documented COVID-19 followed from 365 days before to 455 days after the index (full cohort, 23,790 adults), and those with no Long COVID code by day 90 (landmark cohort, 23,300 adults). We rounded counts up to a multiple of 10 and computed percentages from the rounded counts. –: not shown, because 20 or fewer patients in the group were coded or, for men in the landmark cohort, because we withheld the count so that a reader cannot recover a suppressed count from the cohort total (complementary suppression).

		Full cohort		Landmark cohort (uncoded at day 90)	
		Adults, n	Coded in days 1–455, %	Adults, n	Coded in days 91–455, %
Sex	Female	14,870	3.2	14,550	1.2
	Male	8,730	2.9	8,560	–
	Other/unknown	200	–	200	–
Age, years	18–39	3,840	2.9	3,770	1.3
	40–49	3,200	3.4	3,130	1.3
	50–59	4,360	3.7	4,250	1.2
	60–69	6,070	3.5	5,940	1.3
	70+	6,340	2.7	6,230	1.0
Race and ethnicity	Hispanic or Latino	3,730	2.9	3,660	1.1
	NH Asian	630	–	620	–
	NH Black	2,910	3.1	2,850	1.1
	NH White	14,580	3.2	14,290	1.2
	NH other or multiple	1,280	4.7	1,240	–
	Unknown	680	4.4	660	–

Table 2: Coding fraction by group. Share with a U09.9 code in days 1–455 divided by Household Pulse Survey prevalence among adults who had COVID-19, assuming no false-positive codes and comparable populations. Ratios are relative to the reference group, with 95% intervals reflecting variation across survey periods and in coded shares.

		Coded, %	Ever had Long COVID		Currently has Long COVID	
			Coding fraction	Ratio (95% interval)	Coding fraction	Ratio (95% interval)
Sex	Female	3.2	0.094	0.79 (0.65–0.95)	0.232	0.72 (0.58–0.88)
	Male	2.9	0.120	1 (ref.)	0.322	1 (ref.)
Age, years	18–39	2.9	0.099	0.87 (0.66–1.14)	0.283	1.05 (0.76–1.44)
	40–49	3.4	0.107	0.94 (0.72–1.22)	0.273	1.01 (0.78–1.32)
	50–59	3.7	0.114	1 (ref.)	0.270	1 (ref.)
	60–69	3.5	0.123	1.08 (0.86–1.36)	0.279	1.03 (0.78–1.37)
	70+	2.7	0.115	1.01 (0.74–1.33)	0.253	0.94 (0.67–1.47)
Race and ethnicity	Hispanic or Latino	2.9	0.088	0.79 (0.63–0.98)	0.260	0.98 (0.72–1.31)
	NH White	3.2	0.111	1 (ref.)	0.265	1 (ref.)
	NH Black	3.1	0.104	0.94 (0.73–1.18)	0.290	1.09 (0.82–1.43)
	NH other or multiple	4.7	0.129	1.16 (0.84–1.61)	0.302	1.14 (0.78–1.68)

Table 3: Discrimination of models for a subsequent U09.9 code at day 90. AUROC and AUPRC for the coded outcome, with 95% bootstrap intervals. We pooled the held-out sites.

Model	AUROC (95% CI)	AUPRC (95% CI)
<i>Five-fold cross-validation (23,300 patients)</i>		
M0 prevalence only	0.500	0.011
M1 utilization (logistic)	0.574 (0.537–0.610)	0.015 (0.012–0.019)
M2 elastic net	0.680 (0.643–0.712)	0.034 (0.024–0.048)
M3 gradient boosting	0.681 (0.650–0.709)	0.032 (0.022–0.051)
M4 bagged PU learning	0.663 (0.632–0.695)	0.029 (0.021–0.044)
<i>Held-out sites (18 sites, 5,760 patients)</i>		
M1 utilization (logistic)	0.629 (0.545–0.705)	0.014 (0.009–0.021)
M2 elastic net	0.618 (0.523–0.704)	0.025 (0.013–0.052)
M3 gradient boosting	0.658 (0.579–0.728)	0.030 (0.013–0.077)
M4 bagged PU learning	0.672 (0.588–0.744)	0.028 (0.015–0.060)

Table 4: Exported summary-count models on the validation split. Random 20% of the day-90 cohort (4,680 patients, 50 coded). The application implements naive Bayes with empirical-Bayes shrinkage (main text, Methods).

Model	AUROC (95% CI)	AUPRC (95% CI)
<i>Fit outside the enclave from summary counts (exportable)</i>		
Naive Bayes	0.674 (0.592–0.747)	0.028 (0.016–0.065)
Naive Bayes, empirical-Bayes shrinkage (implemented)	0.674 (0.592–0.747)	0.028 (0.016–0.066)
Linear discriminant, $\lambda = 0.1$	0.667 (0.576–0.749)	0.042 (0.019–0.102)
Linear discriminant, $\lambda = 0.1$, empirical Bayes	0.668 (0.579–0.750)	0.042 (0.019–0.102)
Linear discriminant, $\lambda = 0.3$	0.672 (0.584–0.756)	0.051 (0.019–0.110)
Linear discriminant, $\lambda = 0.3$, empirical Bayes	0.672 (0.582–0.755)	0.051 (0.019–0.111)
Linear discriminant, $\lambda = 0.6$	0.679 (0.598–0.756)	0.036 (0.019–0.091)
Linear discriminant, $\lambda = 0.6$, empirical Bayes	0.680 (0.600–0.755)	0.035 (0.019–0.088)
<i>Fit on participant rows inside the enclave (not exportable)</i>		
Elastic net (M2)	0.680 (0.591–0.757)	0.089 (0.026–0.172)
Bagged PU learning (M4)	0.673 (0.590–0.751)	0.040 (0.020–0.096)

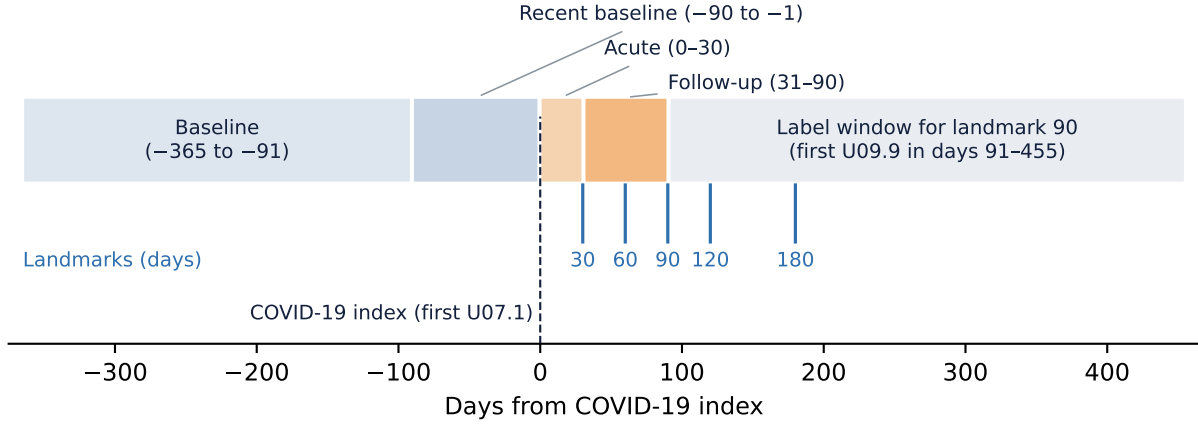


Figure 1: Study design. Feature windows relative to the COVID-19 index date and the label window for the primary landmark (day 90). Blue ticks mark the landmarks at which we rebuilt the cohort, features, and outcome. At landmark L , the follow-up window runs from day 31 to day L and the label window from day $L + 1$ to day 455.

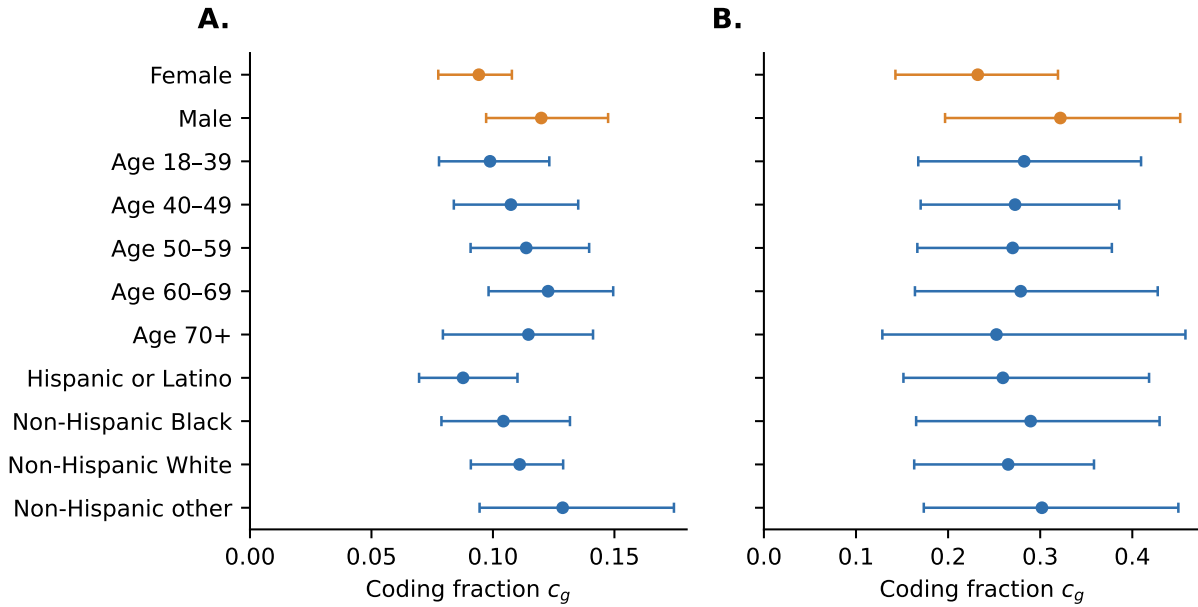


Figure 2: Coding fraction by group. Share of adults with a U09.9 code in days 1–455 divided by Household Pulse Survey prevalence among adults who had COVID-19, with central 95% sensitivity intervals (survey periods and binomial draws). **A.** Prevalence of ever having had Long COVID. **B.** Prevalence of currently having Long COVID. Orange marks the sex groups.

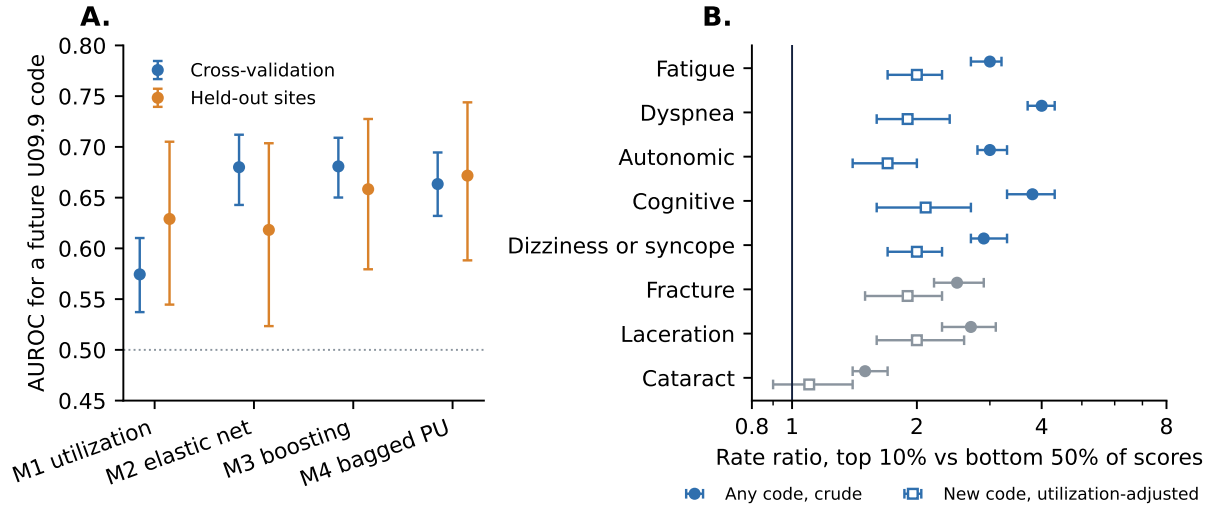


Figure 3: Model performance and specificity at day 90. **A.** AUROC for a future U09.9 code with 95% bootstrap intervals, by five-fold cross-validation and at 18 held-out sites. **B.** Among patients without a subsequent U09.9 code, rate ratios (top 10% vs bottom 50% of elastic-net scores) for codes in days 91–455, for Long COVID–associated symptoms (blue) and negative-control conditions (gray): any code, crude (filled circles), and new code with Mantel–Haenszel adjustment for utilization (open squares), with 95% bootstrap intervals.

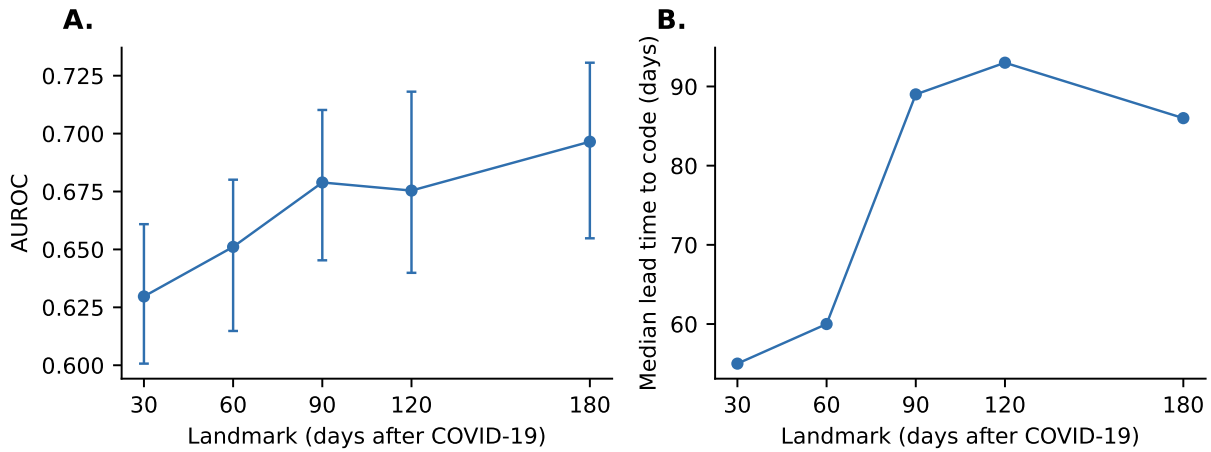


Figure 4: Landmark tasks. We rebuilt the cohort, features, and outcome at each landmark. The outcome is a first U09.9 code from the day after the landmark to day 455, so eligibility and case composition differ by landmark. We restricted landmarks before day 90 to index dates for which every outcome day falls after U09.9 became available. **A.** AUROC with 95% bootstrap intervals for landmark-specific elastic-net models. **B.** Median days from the landmark to the code among later-coded patients in the top 10% of scores.

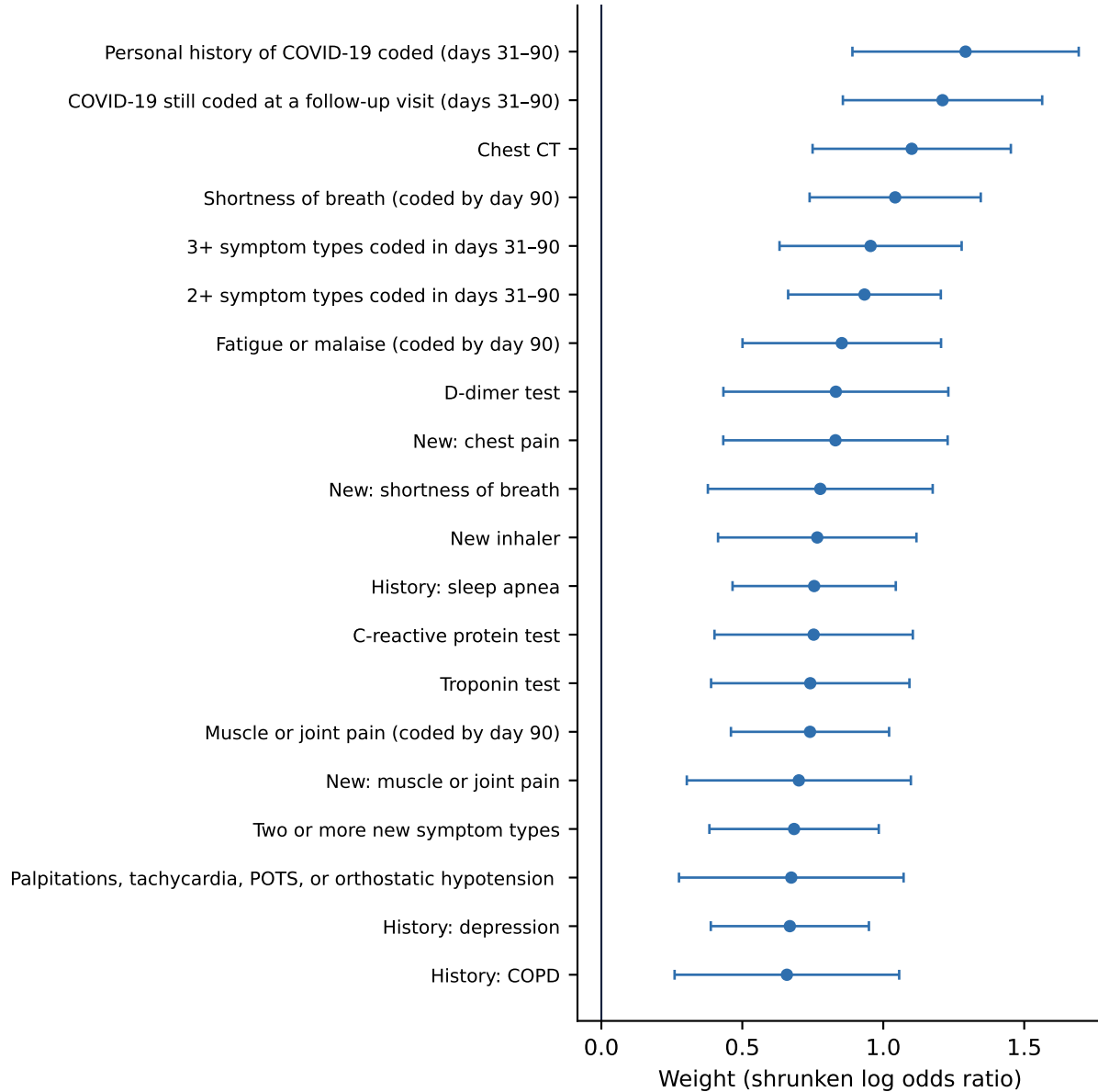


Figure 5: Largest weights in the exported model. The 20 largest empirical-Bayes shrunk log odds ratios of the naive Bayes model fit from released summary counts (training split), with 95% conditional posterior intervals (shrunk weight $\pm 1.96\sqrt{k}$ se, where k is the shrinkage factor, treating the estimated prior variance as known). Each weight is the contribution of a finding to the score when present.