

Supplementary material

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

S1. Cohort selection and decision curve

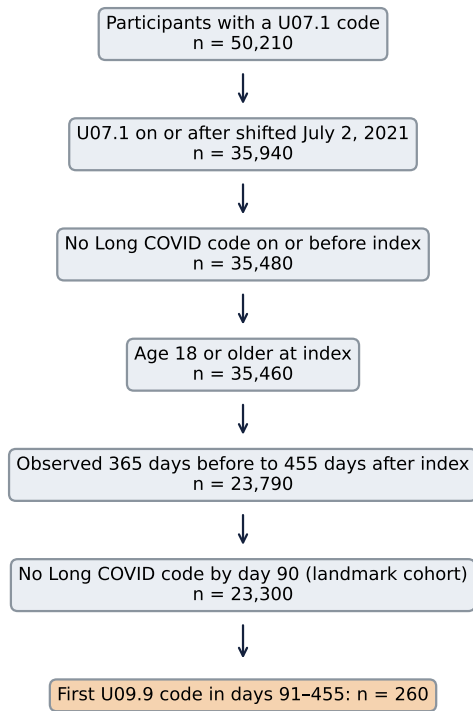


Figure S1: Cohort selection. We rounded counts up to a multiple of 10.

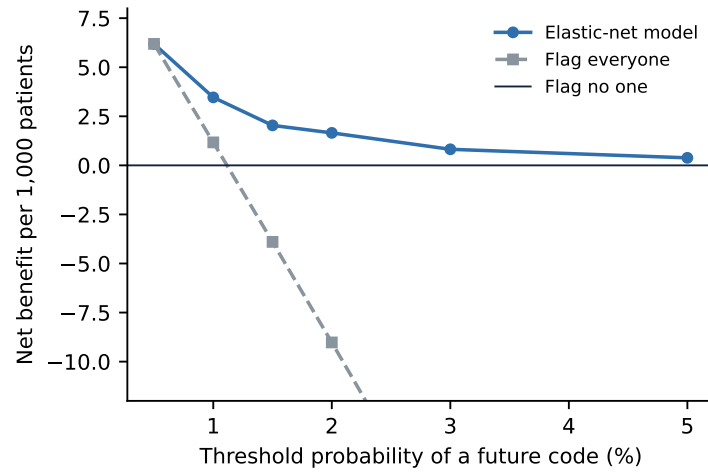


Figure S2: Decision curve for the coded outcome. Net benefit of flagging patients whose predicted probability of a subsequent U09.9 code at day 90 exceeds each threshold, compared with flagging everyone or no one, from the out-of-fold elastic-net scores behind main-text Table 3. We computed net benefit from counts rounded up to a multiple of 10. Uncoded patients with Long COVID count as false positives, so the curve describes the coded endpoint, not detection of disease. Above the 2% threshold, the net benefit of flagging everyone falls below the plotted range.

Table S1: Chart features at the day-90 landmark. Windows: W0 days −365 to −91; W1 days −90 to −1; W2 days 0 to 30; W3 days 31 to 90. “New” features require a match in W2 or W3 and none in W0 or W1. ICD-10-CM codes match by prefix, without dots. Encounter classes: IMP inpatient, EMER emergency, AMB all other visits except pharmacy and laboratory visits. Complete research feature definitions and code lists are provided in the versioned feature registry; the exported model contains the definitions required for its 57 weighted features and 12 auxiliary features.

Feature	Group	Definition	Window
Age 18-39	Demographics	Age at index 18-39 years	at index
Age 40-49	Demographics	Age at index 40-49 years	at index
Age 50-59	Demographics	Age at index 50-59 years	at index
Age 60-69	Demographics	Age at index 60-69 years	at index
Age 70+	Demographics	Age at index 70+ years	at index
Female	Demographics	Female sex at birth	at index
Hospitalized around the COVID diagnosis	Acute illness	Any encounter of class IMP	days −3 to 30
Emergency visit in the first month	Acute illness	Any encounter of class EMER	days 0 to 30
COVID pneumonia	Acute illness	ICD-10-CM J1282	days 0 to 30
Ventilation or intubation	Acute illness	CPT 94002, 94003, 31500; ICD-10-PCS 5A1935Z, 5A1945Z, 5A1955Z	days 0 to 30
4+ outpatient visits in days −365 to −91	Health care use	4 or more encounters of class AMB	W0
2+ outpatient visits in days 31-90	Health care use	2 or more encounters of class AMB	W3
4+ outpatient visits in days 31-90	Health care use	4 or more encounters of class AMB	W3
Emergency visit in days 31-90	Health care use	Any encounter of class EMER	W3
Hospital stay in days 31-90	Health care use	Any encounter of class IMP	W3
Visit rate after COVID at least double the baseline rate	Health care use	Encounters of class AMB or EMER per day in W3 at least 2 times the rate in W0, or any in W3 with none in W0	W3 vs W0
New: fatigue or malaise	New symptoms	ICD-10-CM R53, G933	W2+W3 (none in W0+W1)
New: shortness of breath	New symptoms	ICD-10-CM R060	W2+W3 (none in W0+W1)
New: cough	New symptoms	ICD-10-CM R05	W2+W3 (none in W0+W1)
New: chest pain	New symptoms	ICD-10-CM R07	W2+W3 (none in W0+W1)
New: palpitations, tachycardia, POTS, or orthostatic hypotension	New symptoms	ICD-10-CM R000, R002, I498, G90A, I951	W2+W3 (none in W0+W1)
New: cognitive symptoms (brain fog, memory)	New symptoms	ICD-10-CM R41, G3184	W2+W3 (none in W0+W1)
New: headache	New symptoms	ICD-10-CM R51, G44	W2+W3 (none in W0+W1)
New: dizziness or fainting	New symptoms	ICD-10-CM R42, R55	W2+W3 (none in W0+W1)
New: loss or change of smell or taste	New symptoms	ICD-10-CM R43	W2+W3 (none in W0+W1)

Feature	Group		Definition	Window
New: sleep disturbance	New symptoms		ICD-10-CM G470, G471, G478, G479, F51	W2+W3 (none in W0+W1)
New: anxiety or depression	New symptoms		ICD-10-CM F41, F32, F33, F43	W2+W3 (none in W0+W1)
New: muscle or joint pain	New symptoms		ICD-10-CM M791, M255, M796, M797	W2+W3 (none in W0+W1)
New: gastrointestinal symptoms	New symptoms		ICD-10-CM R10, R11, R14, R19, K58, K59	W2+W3 (none in W0+W1)
New: tingling, numbness, or neuropathy	New symptoms		ICD-10-CM R20, G629	W2+W3 (none in W0+W1)
New: hair loss	New symptoms		ICD-10-CM L659	W2+W3 (none in W0+W1)
New: fever	New symptoms		ICD-10-CM R50	W2+W3 (none in W0+W1)
Two or more new symptom types	New symptoms		2 or more of the 16 new-symptom features above	W2+W3 (none in W0+W1)
Four or more new symptom types	New symptoms		4 or more of the 16 new-symptom features above	W2+W3 (none in W0+W1)
History: asthma	Prior	conditions	ICD-10-CM J45	W0+W1
History: COPD	Prior	conditions	ICD-10-CM J44	W0+W1
History: diabetes	Prior	conditions	ICD-10-CM E10, E11	W0+W1
History: obesity	Prior	conditions	ICD-10-CM E66, Z683, Z684	W0+W1
History: hypertension	Prior	conditions	ICD-10-CM I10	W0+W1
History: depression	Prior	conditions	ICD-10-CM F32, F33	W0+W1
History: anxiety	Prior	conditions	ICD-10-CM F41	W0+W1
History: migraine	Prior	conditions	ICD-10-CM G43	W0+W1
History: fibromyalgia	Prior	conditions	ICD-10-CM M797	W0+W1
History: ME/CFS	Prior	conditions	ICD-10-CM G933	W0+W1
History: chronic pain	Prior	conditions	ICD-10-CM G89	W0+W1
History: sleep apnea	Prior	conditions	ICD-10-CM G4733	W0+W1
History: autoimmune	Prior	conditions	ICD-10-CM M05, M06, M32, L40, K50, K51, M45	W0+W1
History: hypothyroid	Prior	conditions	ICD-10-CM E03	W0+W1
History: heart disease	Prior	conditions	ICD-10-CM I25, I50	W0+W1
Echocardiogram	Tests	after COVID	CPT 93303-93308, 93312-93318, 93350-93352	W2+W3
Holter or event monitor	Tests	after COVID	CPT 93224-93229, 93241-93248, 93268-93272	W2+W3

Feature	Group		Definition	Window
Pulmonary function test	Tests	after	CPT 94010, 94060, 94070, 94726-94729	W2+W3
Chest CT	Tests	after	CPT 71250, 71260, 71270, 71271, 71275	W2+W3
Chest X-ray	Tests	after	CPT 71045-71048	W2+W3
Brain MRI	Tests	after	CPT 70551-70553	W2+W3
Tilt-table test	Tests	after	CPT 93660	W2+W3
Sleep study	Tests	after	CPT/HCPSCS 95806-95811, G0398, G0399, G0400	W2+W3
Six-minute walk test	Tests	after	CPT 94618	W2+W3
D-dimer test	Tests	after	LOINC measurement concepts whose name contains "D-dimer" (12 codes)	W2+W3
Troponin test	Tests	after	LOINC measurement concepts whose name contains "Troponin" (25 codes)	W2+W3
BNP or NT-proBNP test	Tests	after	LOINC measurement concepts whose name contains "Natriuretic peptide" (10 codes)	W2+W3
Thyroid (TSH) test	Tests	after	LOINC measurement concepts whose name contains "Thyrotropin" (98 codes)	W2+W3
Ferritin test	Tests	after	LOINC measurement concepts whose name contains "Ferritin" (9 codes)	W2+W3
C-reactive protein test	Tests	after	LOINC measurement concepts whose name contains "C reactive protein" (13 codes)	W2+W3
Sedimentation rate (ESR) test	Tests	after	LOINC measurement concepts whose name contains "Erythrocyte sedimentation rate" (7 codes)	W2+W3
Vitamin D test	Tests	after	LOINC measurement concepts whose name contains "25-hydroxyvitamin D" (22 codes)	W2+W3
Vitamin B12 test	Tests	after	LOINC measurement concepts whose name contains "Cobalamin" (38 codes)	W2+W3
Antinuclear antibody (ANA) test	Tests	after	LOINC measurement concepts whose name contains "Nuclear Ab" (277 codes)	W2+W3
New beta blocker	New medications		RxNorm ingredients propranolol, metoprolol, atenolol, bisoprolol, nadolol, carvedilol with their descendant drugs (629 concepts)	W2+W3 (none in W0+W1)
New POTS medication	New medications		RxNorm ingredients ivabradine, midodrine, fludrocortisone with their descendant drugs (57 concepts)	W2+W3 (none in W0+W1)

Feature	Group		Definition	Window
New SSRI or SNRI antidepressant	New	medications	RxNorm ingredients sertraline, fluoxetine, citalopram, escitalopram, paroxetine, venlafaxine, duloxetine, desvenlafaxine with their descendant drugs (419 concepts)	W2+W3 (none in W0+W1)
New gabapentin or pregabalin	New	medications	RxNorm ingredients gabapentin, pregabalin with their descendant drugs (142 concepts)	W2+W3 (none in W0+W1)
New inhaler	New	medications	RxNorm ingredients albuterol, fluticasone, budesonide, mometasone, beclomethasone, tiotropium, formoterol, salmeterol with their descendant drugs (1,174 concepts)	W2+W3 (none in W0+W1)
New oral or injected steroid	New	medications	RxNorm ingredients prednisone, methylprednisolone, dexamethasone, prednisolone with their descendant drugs (1,254 concepts)	W2+W3 (none in W0+W1)
New antihistamine or famotidine	New	medications	RxNorm ingredients cetirizine, loratadine, fexofenadine, levocetirizine, famotidine with their descendant drugs (398 concepts)	W2+W3 (none in W0+W1)
New montelukast	New	medications	RxNorm ingredient montelukast with its descendant drugs (24 concepts)	W2+W3 (none in W0+W1)
New naltrexone	New	medications	RxNorm ingredient naltrexone with its descendant drugs (94 concepts)	W2+W3 (none in W0+W1)
New stimulant or wakefulness drug	New	medications	RxNorm ingredients modafinil, armodafinil, methylphenidate, amphetamine, dextroamphetamine, lisdexamfetamine with their descendant drugs (770 concepts)	W2+W3 (none in W0+W1)
New sleep medication	New	medications	RxNorm ingredients zolpidem, trazodone, eszopiclone with their descendant drugs (131 concepts)	W2+W3 (none in W0+W1)
New opioid	New	medications	RxNorm ingredients oxycodone, hydrocodone, tramadol, morphine with their descendant drugs (2,624 concepts)	W2+W3 (none in W0+W1)
Fatigue or malaise (coded by day 90)	Symptoms after COVID		ICD-10-CM R53, G933	days 0 to 90
Shortness of breath (coded by day 90)	Symptoms after COVID		ICD-10-CM R060	days 0 to 90
Cough (coded by day 90)	Symptoms after COVID		ICD-10-CM R05	days 0 to 90
Chest pain (coded by day 90)	Symptoms after COVID		ICD-10-CM R07	days 0 to 90
Palpitations, tachycardia, POTS, or orthostatic hypotension (coded by day 90)	Symptoms after COVID		ICD-10-CM R000, R002, I498, G90A, I951	days 0 to 90

Feature	Group	Definition	Window
Cognitive symptoms (brain fog, memory) (coded by day 90)	Symptoms after COVID	ICD-10-CM R41, G3184	days 0 to 90
Headache (coded by day 90)	Symptoms after COVID	ICD-10-CM R51, G44	days 0 to 90
Dizziness or fainting (coded by day 90)	Symptoms after COVID	ICD-10-CM R42, R55	days 0 to 90
Loss or change of smell or taste (coded by day 90)	Symptoms after COVID	ICD-10-CM R43	days 0 to 90
Sleep disturbance (coded by day 90)	Symptoms after COVID	ICD-10-CM G470, G471, G478, G479, F51	days 0 to 90
Anxiety or depression (coded by day 90)	Symptoms after COVID	ICD-10-CM F41, F32, F33, F43	days 0 to 90
Muscle or joint pain (coded by day 90)	Symptoms after COVID	ICD-10-CM M791, M255, M796, M797	days 0 to 90
Gastrointestinal symptoms (coded by day 90)	Symptoms after COVID	ICD-10-CM R10, R11, R14, R19, K58, K59	days 0 to 90
Tingling, numbness, or neuropathy (coded by day 90)	Symptoms after COVID	ICD-10-CM R20, G629	days 0 to 90
Hair loss (coded by day 90)	Symptoms after COVID	ICD-10-CM L659	days 0 to 90
Fever (coded by day 90)	Symptoms after COVID	ICD-10-CM R50	days 0 to 90
2+ symptom types coded in days 31-90	Symptoms after COVID	2 or more of the 16 symptom clusters coded (new or not)	days 31 to 90
3+ symptom types coded in days 31-90	Symptoms after COVID	3 or more of the 16 symptom clusters coded (new or not)	days 31 to 90
COVID-19 still coded at a follow-up visit (days 31-90)	Coding after COVID	ICD-10-CM U071	days 31 to 90
Sequelae of infection coded after COVID	Coding after COVID	ICD-10-CM B948	days 0 to 90
Personal history of COVID-19 coded (days 31-90)	Coding after COVID	ICD-10-CM Z8616	days 31 to 90
Visit in days 61-90	Health care use	Any encounter of class AMB or EMER	days 61 to 90

S2. Sensitivity of the coding fraction

The primary coding fraction assumes no false-positive codes ($q_g = 1$) and equal prevalence in the All of Us cohort and the survey population ($\kappa_g = 1$). Under Equation 2 of the main text, the women-to-men coding ratio equals the primary ratio multiplied by $(q_F/q_M)/(\kappa_F/\kappa_M)$. Figure S3 shows the ratio over a grid of both quantities. The point estimate reaches 1 only when $(q_F/q_M)/(\kappa_F/\kappa_M)$ is at least 1.27 (ever definition) or 1.39 (current definition). The upper interval bound reaches 1 at 1.05 or 1.13. When we added each survey period’s sampling error to the bootstrap, the interval for women widened from 0.65–0.95 to 0.65–0.96 (ever) and from 0.58–0.88 to 0.56–0.92 (current). For composition reweighting, we used nine HPS public-use waves (weeks 46, 48, 50, 51, 54, 56, 58, 60, and 62). Their unweighted-composition prevalence by sex reproduced the published series for the same weeks to within 0.1 percentage point. After we raked within each sex to the All of Us cohort’s age-band and race-and-ethnicity margins (computed from counts rounded up to a multiple of 10), mean prevalence changed by factors of 0.98 (women) and 0.97 (men) for ever having Long COVID and 1.03 and 1.09 for current Long COVID. This adjusts composition only. Differences in health and health-care use between volunteers and survey respondents, and between EHR-documented and self-reported COVID-19, remain in κ_g . To gauge the size of a sex difference in κ , we estimated it for two conditions that are coded reliably and measured in a national survey. Among adults with at least one EHR visit in each of two consecutive years who completed the All of Us Personal and Family Health History survey, the share with any ICD-10-CM code among participants reporting a clinician’s diagnosis was 0.78 (women) and 0.82 (men) for hypertension and 0.81 and 0.88 for type 2 diabetes. We standardized each sex’s coded share to the age distribution of US adults (NHANES 2021–2023 interview weights) and divided it by that share to estimate the cohort’s diagnosed prevalence. Dividing this prevalence by the prevalence of an ever-told diagnosis in the CDC/NCHS 2023 National Health Interview Survey (Sample Adult public-use file, survey weights) gives the cohort-to-national prevalence ratio. The ratios were 1.43 (women) and 1.49 (men) for hypertension and 2.42 and 2.64 for type 2 diabetes, so κ_F/κ_M was 0.96 and 0.92 (0.94 and 0.89 when also standardized to race and ethnicity). These conditions are not Long COVID, and κ differed by condition, so the values indicate the plausible size of a sex difference rather than its value for Long COVID. With $q_F = q_M$, the women-to-men coding ratio reaches 1 only if κ_F/κ_M is 0.79 or lower (ever) or 0.72 or lower (current). When we changed the weights that combine HPS age groups (0.3 to 0.7 for ages 18–29 within 18–39, and 0.5 to 0.8 for 70–79 within 70 or older), age-band coding ratios changed by 0.03 or less.

Index dates before U09.9 was available. For index dates from July 2 to September 29, 2021 (shifted), some of days 1 to 90 can fall before U09.9 was introduced on October 1, 2021, which could lower the coded share in the coding-fraction cohort. When we restricted the cohort to shifted index dates on or after September 30, 2021, it had 18,830 adults, and their coded share in days 1 to 455 was 2.8%. The restriction lowered the coded share among women, in every age band, and in every racial and ethnic group shown except non-Hispanic Black adults, whose share was similar: adults with earlier index dates were coded more often, not less, so including them did not lower the estimated coding fractions.

S3. Related post-infectious conditions

ME/CFS and postviral fatigue (G93.3x), POTS and dysautonomia (G90.A, G90.9), fibromyalgia (M79.7), mast cell disorders (D89.4x), and chronic fatigue (R53.82) share symptoms and coding barriers with Long COVID. At day 90, incident codes in days 91 to 455 numbered 80 for ME/CFS

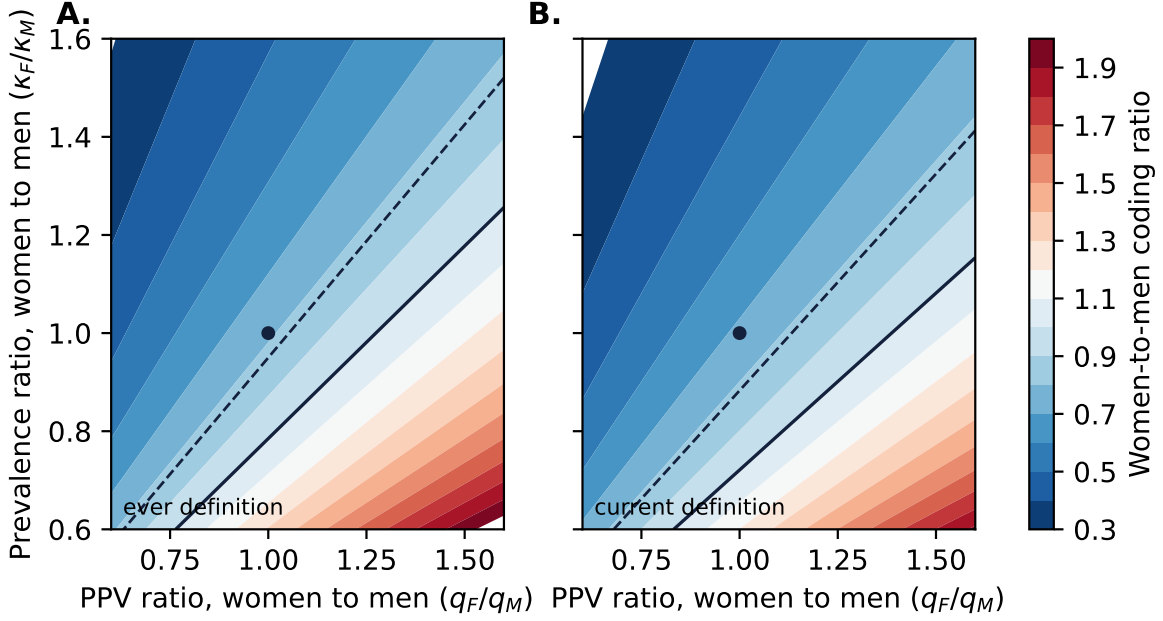


Figure S3: Women-to-men coding ratio under differences in code validity and survey comparability. Shading shows the ratio as the women-to-men ratio of positive predictive values (q_F/q_M) and of cohort-to-survey prevalence (κ_F/κ_M) vary. The solid line marks where the point estimate equals 1, the dashed line where the upper 95% bound equals 1, and the dot the primary analysis. **A.** Ever-had definition. **B.** Current definition.

or postviral fatigue (30 of them also coded with Long COVID), 80 for POTS or dysautonomia, 210 for fibromyalgia, 240 for chronic fatigue, and 20 or fewer for mast cell disorders, which we excluded. We compared a single-task logistic model for Long COVID with a multi-task logistic model in which each condition k has weights $\beta_k = \beta_0 + \delta_k$, with the deviations δ_k penalized toward the shared weights, and with a model of a pooled outcome (Long COVID or any related code). We fit the standardization inside each training fold, and each condition had its own intercept. With identical folds, AUROC for Long COVID was 0.662 (0.632–0.697) for the single-task model, 0.672 (0.643–0.702) for the multi-task model, and 0.666 (0.632–0.699) for the pooled outcome. Borrowing strength from related conditions changed discrimination little.

S4. Subgroups and analysis provenance

Table S2: Discrimination, calibration, and flagging of the day-90 elastic net by group. Out-of-fold predictions from five-fold cross-validation with folds different from those of main-text Table 3, so the overall AUROC differs slightly from the primary 0.680. We rounded counts up to a multiple of 10 and computed the observed rate from the rounded counts, so observed-to-predicted ratios are approximate. Flagged: share of the group among the top 5% of all scores. We do not show groups with 20 or fewer coded patients, whose estimates are unstable. –: withheld for men, because with the overall row these values would let a reader recover the suppressed values for other or unknown sex (complementary suppression).

	Group	Patients	Coded	AUROC (95% CI)	Observed/predicted	Flagged at 5% workload, %
All patients		23,300	260	0.688 (0.650–0.724)	1.01	5.0
Sex	Female	14,550	170	0.697 (0.657–0.739)	1.03	5.4
	Male	8,560	–	0.667 (0.603–0.719)	–	–
Race and ethnicity	Hispanic or Latino	3,660	40	0.676 (0.590–0.747)	1.02	4.4
	NH Black	2,850	30	0.697 (0.569–0.808)	0.85	7.0
	NH White	14,290	170	0.700 (0.658–0.739)	1.11	4.8
Age, years	18–39	3,770	50	0.759 (0.675–0.837)	1.27	4.5
	40–49	3,130	40	0.541 (0.446–0.622)	1.14	5.1
	50–59	4,250	50	0.657 (0.576–0.733)	1.02	5.6
	60–69	5,940	80	0.733 (0.674–0.781)	1.16	5.7
	70+	6,230	60	0.685 (0.616–0.756)	0.94	4.5

Table S3: Exploratory score adjustment. We divided the day-90 elastic-net out-of-fold scores of women by the women-to-men coding ratio r (men and other sex unchanged) and computed shares from counts rounded up to a multiple of 10.

Ranking	Definition	Ratio r	Women in top 5%, %	Recall of coded in top 5%, %	AUROC
Unadjusted	–	1.00	67	19	0.688
Adjusted, point ratio	ever	0.79	75	23	0.684
Adjusted, lower bound	ever	0.65	80	19	0.678
Adjusted, upper bound	ever	0.95	69	19	0.688
Adjusted, point ratio	current	0.72	78	19	0.681
Adjusted, lower bound	current	0.58	84	19	0.673
Adjusted, upper bound	current	0.88	72	19	0.687
Cohort	–	–	62	–	–

Table S4: Cohort, features, evaluation, and model behind each result. “Matrix A” is the day-90 matrix of 102 features that we built once for the main analysis, with ICD-10-CM codes from the condition and observation tables. “Rebuilt” matrices use the same code sources, and we recomputed them at each landmark from the same registry with windows ending at the landmark (we defined the recent-symptom and late-visit windows relative to the landmark), so day-90 estimates differ slightly between the two. B: bootstrap replicates.

Result	Cohort and outcome	Features	Evaluation	Model
Coding fractions (Table 2, Figure 2, Section S2)	23,790 adults, U09.9 in days 1–455	none	survey periods and binomial draws, B = 5,000	none
Day-90 prediction, lead time, documentation checks (Table 3, Figure 3)	23,300 uncoded at day 90, first U09.9 in days 91–455	matrix A	5-fold stratified CV (B = 200), 18 held-out sites pooled, and documentation ratios (B = 200)	M0–M4
Landmark tasks (Figure 4)	rebuilt at each landmark (early landmarks restricted by index date), first U09.9 in days $L + 1$ –455	rebuilt	5-fold stratified CV, B = 200	elastic net
Day-30 vs day-180 scores (Results)	eligible at day 180 and in the day-30 cohort, first U09.9 in days 181–455	rebuilt	out-of-fold scores of two separately trained models, B = 200	elastic net
Subgroups (Table S2)	as day-90 prediction	matrix A	5-fold stratified CV (folds different from Table 3), B = 200	elastic net
Score adjustment, workload comparison (Table S3, Results)	as day-90 prediction	matrix A	5-fold stratified CV (folds of Table S2)	elastic net, M1, symptom-count rule
Decision curve (Figure S2)	as day-90 prediction	matrix A	out-of-fold scores of Table 3	elastic net
Related conditions (Section S3)	landmark 90, incident related codes	rebuilt	5-fold stratified CV (identical folds), B = 200	logistic, multi-task
Exported model (Table 4)	as day-90 prediction, split at random into 80% training and 20% validation (seed 20260929)	matrix A, 57 eligible features	validation split, B = 200	naive Bayes, linear discriminants
Exported pipeline at held-out sites (Results)	18 held-out sites, with training sites split 70/30 into fit and calibration	matrix A, 41 eligible features	held-out sites, paired B = 500	naive Bayes (shrinkage), elastic net
Utilization before documentation (Results)	later-coded patients vs uncoded in matched strata	visit and test counts	bootstrap	none

Result	Cohort and outcome	Features	Evaluation	Model
Binary vs richer features (Section S8)	23,300 uncoded at day 90, first U09.9 in days 91–455	matrix A and rich counts	5-fold stratified CV and 18 held-out sites, B = 200	elastic nets, gradient boosting
Era-safe coded share (Results, Section S2)	18,830 adults with shifted index on or after September 30, 2021, U09.9 in days 1–455	none	none	none
Laboratory results and vital signs (Results, Section S6)	as day-90 prediction	matrix A plus laboratory-result and vital-sign features	5-fold stratified CV (folds of Table S2) and 18 held-out sites, B = 200	elastic net
Alternative model families (Section S7, Table S6)	as day-90 prediction	matrix A	nested 5-fold CV and 18 held-out sites, B = 200	boosted trees, random forest, decision tree, elastic net
Development and evaluation populations, source grouping (Tables S8 and S9)	as day-90 prediction, both splits	selected features of matrix A	none	none

S5. Coded-outcome and disease discrimination

Within one eligible population at landmark L , with coded share p_L , disease prevalence π_L , coding fraction $c_L = p_L/\pi_L$, no false-positive codes, and coding that does not depend on the features among people with the disease, the share of the uncoded who have the disease is $\alpha_L = (\pi_L - p_L)/(1 - p_L)$, and $\text{AUC}_{\text{coded}} = (1 - \alpha_L) \text{AUC}_{\text{disease}} + \alpha_L/2$, as in Jain, White, and Radivojac (2017) and Ramola, Jain, and Radivojac (2019). The survey does not identify π_L for patients still uncoded at day 90, so Table S5 is hypothetical: with $p_{90} = 0.011$ and the cross-validated coded-outcome AUROC of 0.680, the implied disease AUROC is 0.682 if $\pi_{90} = 0.02$, 0.698 if $\pi_{90} = 0.10$, 0.723 if $\pi_{90} = 0.20$, and 0.754 if $\pi_{90} = 0.30$.

Table S5: Implied disease AUROC under hypothetical disease prevalence among patients uncoded at day 90.

π_{90}	α_{90}	Implied disease AUROC
0.02	0.009	0.682
0.05	0.039	0.687
0.10	0.090	0.698
0.15	0.141	0.709
0.20	0.191	0.723
0.30	0.292	0.754

S6. Supplementary methods

Exportable candidates. From each feature’s rounded 2×2 table with the outcome (minus the expected rounding excess of 4.5 per cell), naive Bayes uses weights equal to the log odds ratio w_j . The empirical-Bayes version replaces w_j by $\tilde{w}_j = w_j \tau^2 / (\tau^2 + \text{se}_j^2)$, with se_j^2 the sum of reciprocal cell counts and $\tau^2 = \max\{0, \overline{w^2} - \overline{\text{se}^2}\}$. Linear discriminants use $w = \Sigma^{-1}(\mu_1 - \mu_0)$, where class means come from the label tables and the within-class covariance Σ_W from released pair counts (independence values where we suppressed a pair). We shrank Σ_W as $(1 - \lambda)\Sigma_W + \lambda \text{diag}(\Sigma_W)$ for $\lambda \in \{0.1, 0.3, 0.6\}$ and fit each version with and without empirical-Bayes shrinkage of the standardized mean differences.

Rounding perturbation. Across 200 replicates that replaced each rounded count independently by a compatible value, refit naive Bayes weights correlated at least 0.95 with the published weights (median 0.97), and the synthetic high-risk patient stayed in the High band, with the central 95% of its predicted probabilities at about 5.5% to 6.3%. We perturbed counts independently, without enforcing joint compatibility. The within-class covariance matrix built from the counts can be slightly indefinite (smallest eigenvalue across replicates -0.0004), which shrinkage toward the diagonal mitigates.

Laboratory results and vital signs. We counted a laboratory result in days 0 to 90 as abnormal if it lay outside the recorded reference range in the direction that applies to the test (above the upper bound for most tests, below the lower bound for vitamin D and B12, outside either bound for TSH), if its value concept indicated an abnormal result (for example, high, low, abnormal, positive, detected, or reactive), or, when the bound for that direction was missing, beyond unit-specific thresholds (for example, D-dimer above $0.5 \mu\text{g/mL}$ FEU, C-reactive protein above 10 mg/L , ferritin above 300 ng/mL). We read value concepts with negation-aware word rules: a flag word preceded by “not,” “non,” “no,” “none,” or “without,” or followed by “normal,” did not count, so “non-reactive” and “high normal” were not abnormal. Each direction used its own recorded bound, and we matched units on their UCUM codes as well as their names. Vital-sign features were any heart rate above 100 per minute in days 31 to 90, a mean heart rate in days 31 to 90 above 90 per minute or at least 10 per minute above the baseline mean, and any oxygen saturation below 94% in days 0 to 90. Each laboratory test had a result in days 0 to 90 for 1% to 16% of patients, and 16% to 89% of tested participants had at least one result with a recorded upper reference bound. When we added the corrected laboratory-result and vital-sign features to the 102 features of matrix A, the cross-validated AUROC was 0.682 (95% CI 0.643–0.719), against 0.688 (0.651–0.725) without them (a paired difference of -0.006 , 95% CI -0.012 to 0.001), and the held-out-site AUROC was 0.624 (0.546–0.715) against 0.618 (0.544–0.722) (a difference of 0.006 , 95% CI -0.012 to 0.026).

Survey reweighting. HPS reports ages 18–29 and 30–39, and 70–79 and 80 or older, separately. In the primary analysis, we combined them with fixed weights of 0.48 and 0.52, and 0.65 and 0.35. For composition reweighting, we raked HPS person weights among adults who had COVID-19, separately for women and men, to the All of Us cohort’s age-band and race-and-ethnicity margins for that sex.

S7. Alternative model families

We tuned gradient-boosted trees by an inner three-fold cross-validation over tree depth (2 to 4), learning rate (0.03 or 0.1), and number of trees (200 or 400), nested inside the outer folds so that tuning never saw the evaluation fold. XGBoost and LightGBM were not available in the Researcher Workbench environment, so we used scikit-learn’s histogram-based implementation. The random forest had 500 trees with at least 20 patients per leaf, and the decision tree had depth 4 with at least 100 patients per leaf. These analyses used the feature matrix, outcome, and evaluation partitions of Table 3.

Table S6: Alternative model families for a subsequent U09.9 code at day 90. 95% bootstrap intervals (200 replicates). We pooled the held-out sites (5,760 patients).

Model	AUROC (95% CI)	AUPRC
<i>Five-fold cross-validation</i>		
random forest	0.672 (0.642–0.698)	0.023
decision tree (depth 4)	0.614 (0.582–0.650)	0.019
elastic net (M2)	0.680 (0.651–0.710)	0.034
tuned boosted trees (scikit-learn, nested tuning)	0.648 (0.616–0.680)	0.023
<i>Held-out sites</i>		
random forest	0.665 (0.588–0.734)	0.025
decision tree (depth 4)	0.619 (0.538–0.698)	0.017
elastic net (M2)	0.618 (0.523–0.713)	0.025
tuned boosted trees (scikit-learn, nested tuning)	0.624 (0.524–0.706)	0.024

S8. Exploratory comparison with richer features

We built the exported model from 2×2 tables, so every feature must be binary. We compared the export-compatible binary feature set with an exploratory feature set containing continuous age and event counts, refitting participant-level models on the day-90 landmark cohort with these richer features: continuous age and its square, the number of symptom clusters coded in days 31 to 90, and log counts of visits by class (outpatient, emergency, inpatient) and window, of distinct coding dates for every condition group before and after the index, and of distinct dates of each test, laboratory order, and prescription. Condition codes came from both the condition and observation tables. We fit an elastic net (ℓ_1 ratio 0.5, $C = 0.05$) on the binary features (reproducing M2), on the rich features, and on both. We also fit an elastic net on the rich features with C chosen by inner three-fold cross-validation from 0.01, 0.05, 0.2, and 1, and scikit-learn histogram gradient boosting on the rich features (300 iterations, learning rate 0.05, 15 leaves, at least 40 patients per leaf). Folds and held-out sites were those of the main analysis. The separately fitted summary-count score reached 0.659 (0.576–0.722) at the held-out sites (main text).

Table S7: Binary and richer features for a subsequent U09.9 code at day 90. 95% bootstrap intervals (200 replicates). We pooled the held-out sites (5,760 patients).

Model	Features	AUROC (95% CI)	AUPRC
<i>Five-fold cross-validation</i>			
elastic net, binary features (as M2)	102	0.680 (0.651–0.710)	0.034
elastic net, rich features	123	0.692 (0.659–0.723)	0.033
elastic net, binary plus rich features	225	0.705 (0.674–0.737)	0.042
elastic net, rich features, tuned penalty (C 0.01 or 0.05)	123	0.691 (0.660–0.728)	0.029
gradient boosting, rich features	123	0.694 (0.666–0.723)	0.034
<i>Held-out sites</i>			
elastic net, binary features (as M2)	102	0.618 (0.523–0.713)	0.025
elastic net, rich features	123	0.600 (0.520–0.678)	0.016
elastic net, binary plus rich features	225	0.617 (0.524–0.702)	0.022
elastic net, rich features, tuned penalty (C 0.01 or 0.05)	123	0.673 (0.594–0.750)	0.019
gradient boosting, rich features	123	0.708 (0.647–0.764)	0.019

S9. Development and evaluation populations and source grouping

Table S8 compares the populations that we used to develop and evaluate the models: the random training and validation splits of the exported-model analysis, and the held-out and training sites of the site evaluation. Table S9 describes the source identifiers behind the site grouping. We defined sites by the source of the index event. Records in the feature tables could come from other sources, so the held-out sites were not fully separated from the training sites.

Table S8: Development and evaluation populations at day 90. The training split is the random 80% of the landmark cohort that we used to fit the exported model, and the validation split is the other 20%. The held-out sites are the 18 sites that we excluded from fitting in the site evaluation, and the training sites are all other sites. Each pair of columns partitions the landmark cohort. Values are n (%). We rounded counts up to a multiple of 10 and computed percentages from the rounded counts. ≤ 20 : 20 or fewer patients. –: we withheld the count so that a reader cannot recover the suppressed count from the other cells (complementary suppression).

	Random split		Site split	
	Training split	Validation split	Held-out sites	Training sites
Patients, n	18,620	4,680	5,760	17,540
First U09.9 code in days 91–455	210 (1.1)	50 (1.1)	60 (1.0)	210 (1.2)
<i>Sex</i>				
Female	11,640 (62.5)	2,920 (62.4)	4,000 (69.4)	10,560 (60.2)
Male	6,830 (36.7)	1,730 (37.0)	1,710 (29.7)	6,850 (39.1)
Other/unknown	160 (0.9)	40 (0.9)	60 (1.0)	140 (0.8)
<i>Age, years</i>				
18–39	2,990 (16.1)	780 (16.7)	1,050 (18.2)	2,720 (15.5)
40–49	2,530 (13.6)	600 (12.8)	810 (14.1)	2,330 (13.3)
50–59	3,410 (18.3)	850 (18.2)	1,060 (18.4)	3,200 (18.2)
60–69	4,730 (25.4)	1,210 (25.9)	1,470 (25.5)	4,470 (25.5)
70+	4,990 (26.8)	1,250 (26.7)	1,380 (24.0)	4,850 (27.7)
<i>Race and ethnicity</i>				
Hispanic or Latino	2,920 (15.7)	740 (15.8)	350 (6.1)	3,310 (18.9)
NH Asian	490 (2.6)	140 (3.0)	90 (1.6)	540 (3.1)
NH Black	2,290 (12.3)	560 (12.0)	710 (12.3)	2,140 (12.2)
NH White	11,410 (61.3)	2,890 (61.8)	4,240 (73.6)	10,060 (57.4)
NH other or multiple	1,000 (5.4)	250 (5.3)	240 (4.2)	1,000 (5.7)
Unknown	540 (2.9)	130 (2.8)	150 (2.6)	520 (3.0)
<i>Chart features by day 90</i>				
Hospitalized around the COVID-19 diagnosis	2,590 (13.9)	600 (12.8)	700 (12.2)	2,490 (14.2)
Emergency visit in days 0 to 30	2,830 (15.2)	680 (14.5)	930 (16.1)	2,580 (14.7)
4+ outpatient visits in days –365 to –91	14,490 (77.8)	3,630 (77.6)	4,600 (79.9)	13,510 (77.0)
2+ outpatient visits in days 31 to 90	11,710 (62.9)	2,920 (62.4)	3,470 (60.2)	11,160 (63.6)
Two or more new symptom types	2,770 (14.9)	700 (15.0)	760 (13.2)	2,710 (15.5)
COVID-19 still coded in days 31 to 90	1,050 (5.6)	290 (6.2)	200 (3.5)	1,130 (6.4)
History of obesity	5,210 (28.0)	1,300 (27.8)	1,500 (26.0)	5,010 (28.6)
History of hypertension	8,630 (46.3)	2,090 (44.7)	2,460 (42.7)	8,250 (47.0)
History of depression	4,310 (23.1)	1,050 (22.4)	1,460 (25.3)	3,910 (22.3)

Table S9: Source identifiers and records from other sources. Day-90 landmark cohort. We assigned each participant the source identifier of the index event. The participant-mediated EHR sources are PTSC (Participant Technology Systems Center) and TPC (The Participant Center). We counted a patient when they had at least one record of the kind described in days -365 to 90 . The lower panel gives the share of each feature table’s records in days -365 to 90 that came from a source other than the participant’s index-event source (no record in any table lacked a source identifier). We rounded counts up to a multiple of 10 and computed percentages from the rounded counts.

Measure	Patients, n (%)
<i>Landmark cohort (23,300 patients)</i>	
Index-event source type: EHR site	21,770 (93.4)
Index-event source type: participant-mediated EHR, PTSC	1,100 (4.7)
Index-event source type: participant-mediated EHR, TPC	440 (1.9)
Index-event source identifier missing	0 (0.0)
Index day carries U07.1 records from more than one source	220 (0.9)
Any record in days -365 to 90 from a source other than the index-event source	12,430 (53.3)
Any record in days -365 to 90 with a missing source identifier	0 (0.0)
No record in days -365 to 90 in the feature tables	0 (0.0)
Distinct index-event source identifiers	54
Distinct source identifiers on feature-table records	67
<i>Held-out sites (5,760 patients)</i>	
Any record in days -365 to 90 from a training-site source	710 (12.3)
<i>Training sites (17,540 patients)</i>	
Any record in days -365 to 90 from a held-out-site source	100 (0.6)
Records in days -365 to 90 from a source other than the index-event source, by feature table	Records, %
Condition	6.5
Observation	46.3
Visit	9.5
Procedure	9.2
Measurement	9.5
Drug exposure	15.0