

Design and Technical Evaluation of a Standards-Based System That Delivers Enclave-Derived Case-Finding Models to Patients and Clinicians

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

Background. Research data enclaves such as the All of Us Research Program do not allow participant-level data, or models fitted to those data, to leave the enclave, which makes it hard to deliver case-finding models for undercoded conditions to patients and clinicians.

Objectives. We describe and technically evaluate a system that fits models outside an enclave from released summary counts and delivers them through standard health information technology interfaces.

Methods. We exported dissemination-compliant summary counts from the All of Us Researcher Workbench and, outside the enclave, fitted versioned models for Long COVID, chronic fatigue, fibromyalgia, endometriosis, Lyme disease, lupus, migraine, depression, and anxiety. A written model contract, a Python reference implementation, and synthetic Fast Healthcare Interoperability Resources (FHIR) test patients define the expected outputs. One JavaScript scorer runs in a patient demonstration, a SMART on FHIR clinician application, and a CDS Hooks service. An atlas displays aggregate results. We evaluated the system on synthetic records with automated tests, cross-implementation output comparison, and AI-generated adversarial checks.

Results. The deployed Long COVID model, one of the summary-count models compared in a companion analysis, had an area under the receiver operating characteristic curve of 0.67 in a random validation split. A separately fitted version of the procedure reached 0.66 at 18 held-out sites. The chronic fatigue, fibromyalgia, Lyme disease, and lupus models reached 0.74, 0.73, 0.65, and 0.82 for self-reporters without a code. The JavaScript scorer reproduced all 180 reference outputs exactly, and 457 of 474 automated tests passed (17 skipped). Adversarial checks found 5 classes of failure, including scoring of ineligible patients and treating unavailable

data as absent. We fixed all 5 and added them as tests.

Conclusions. A model contract fitted from released summary counts can be delivered through FHIR, SMART on FHIR, and CDS Hooks with reproducible scoring on synthetic records. Real-record interoperability, usability, and prospective clinical performance remain to be evaluated.

Keywords: clinical decision support; SMART on FHIR; CDS Hooks; privacy-preserving model sharing; electronic health records

1 Background and Significance

Conditions such as Long COVID, myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS), fibromyalgia, and Lyme disease are often diagnosed late or not recorded with a diagnosis code. In companion analyses of the All of Us Research Program, the ratio of recorded Long COVID diagnoses to survey-based prevalence was about 1 in 9, under assumptions about comparable populations [1]. Among respondents reporting ongoing care for chronic fatigue with at least 5 years of electronic health record (EHR) history, the coded share was 0.28 times the mean coded share for hypertension and type 2 diabetes [2]. Case-finding models could help surface these patients, but the data needed to build them often sit in research enclaves.

The All of Us Researcher Workbench allows only aggregate results that meet its dissemination policy to leave the enclave, and it does not allow models trained on participant rows to be exported [3]. These restrictions create an implementation problem: a deployable model must be constructed without exporting participant-level data or a model fitted directly to them. Separately, published models are often not deployed, because deployment requires engineering for data access, scoring, and display inside clinical systems [4]. Standards exist for each step: FHIR R4 for data [5], SMART on FHIR for launching applications from the EHR [6, 7], and CDS Hooks for returning decision-support cards to the EHR [8].

2 Objectives

We describe the design of a system that (1) fits deployable models outside the enclave from released summary counts, (2) runs them through one scoring implementation in patient, clinician, and integration settings, and (3) uses FHIR R4, SMART on FHIR, and CDS Hooks so that compatible

EHR implementations could adopt it. We report a technical evaluation on synthetic records, organized following STARE-HI where it applies [9].

3 Methods

3.1 Design requirements

The design followed 5 requirements: no participant-level data or participant-trained model leaves the enclave, one model definition drives every application, every clinician-facing output shows the chart findings behind it, patient-facing screens show no model probabilities or risk multiples, and public demonstrations use synthetic patients only. Figure 1 shows the architecture and data boundaries.

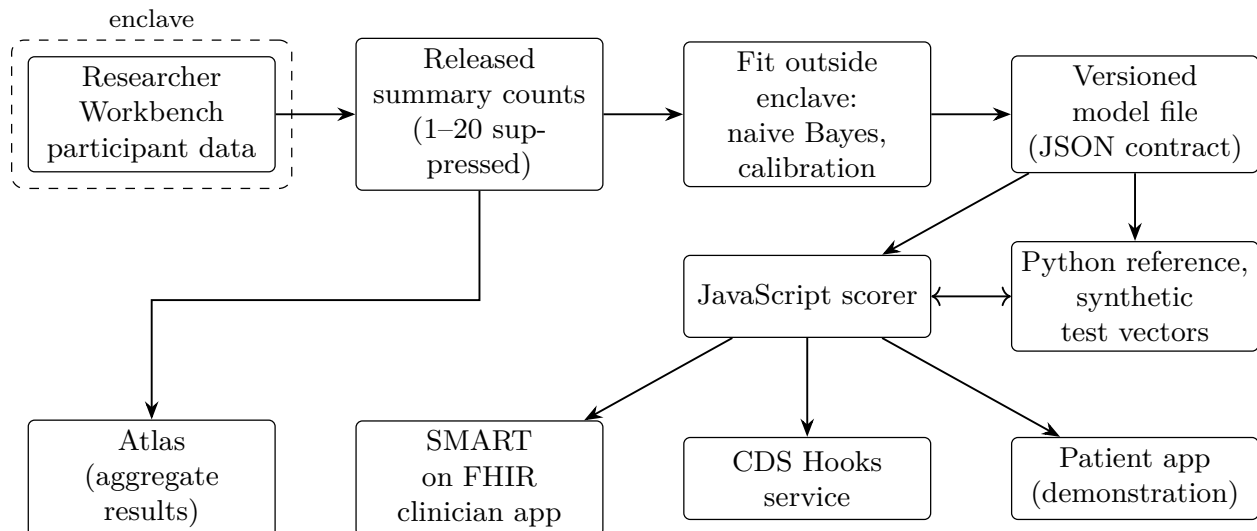


Figure 1: System architecture and data boundaries. Only summary counts that pass the dissemination checks leave the All of Us Researcher Workbench (dashed boundary). We fit models outside the enclave and store them as versioned JSON files. A Python reference implementation and synthetic FHIR test patients define the expected outputs, which the JavaScript scorer must reproduce (double arrow). The scorer runs in the clinician application, the CDS Hooks service, and the patient demonstration. The atlas shows aggregate results only.

3.2 Models fitted from released counts

The underlying analyses used the All of Us Research Program Registered Tier Dataset v9 (curated data repository R2025Q4R6, data cutoff January 1, 2025) [10]. Inside the Workbench, 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 suppressed count could otherwise be recovered from the table’s totals (complementary suppression). We retained features only when every cell of their feature-by-label table exceeded 20 [3]. Outside the enclave, we fitted each model as naive Bayes with empirical-Bayes shrinkage of each feature’s log odds ratio and calibrated it from released score bins [1, 11]. The Long COVID model predicts subsequent documentation: day 0 is the first recorded COVID-19 diagnosis (ICD-10-CM U07.1) on or after July 2, 2021, as in the training cohort, predictors use records through day 90, and the outcome is a first U09.9 code in days 91 to 455. The chronic fatigue, fibromyalgia, Lyme disease, and lupus models predict a self-reported prior diagnosis among adults with no qualifying code (systemic lupus erythematosus [M32] for lupus, A69.2 for Lyme disease), from records in the 5 years before the scoring date. The Lyme disease and lupus models add condition-specific findings, tests, and treatments to the same feature set. The chronic fatigue label does not establish ME/CFS, the Lyme disease label mostly reflects past, treated infection, and we have not validated any self-report model against an independent assessment of current disease. Risk bands (Typical, Elevated, High) are heuristic display thresholds at 1.5 and 3 times the development-cohort label prevalence.

3.3 Model contract and test vectors

A written specification defines how to evaluate a model file on a FHIR R4 bundle: which resources and codes count for each feature, where windows fall relative to the index event or the scoring date, when not to score a patient, how nested count thresholds collapse to one finding, and how the score maps to a probability, a band, and a multiple of the average. The scorers exclude a patient when the birth date is missing, when the patient is younger than 18 years on day 0 (the index for Long COVID and the scoring date for the other models), when a code for the target condition is recorded by the scoring date or is undated, or, for Long COVID, when fewer than 90 or more than 455 days have passed since the index. Only records dated on or before the scoring date count. The scorers ignore refuted conditions and entered-in-error records and accept partial FHIR dates, reading a partial record date as the earliest compatible day and a partial birth date as July 1 or the 15th of the month. A failed, forbidden, or truncated FHIR search yields an explicit could-not-score result instead of an absent finding. The Python reference implementation defines the expected software outputs. Eighteen synthetic FHIR patients, each built to exercise specific rules (10 demonstration

patients and 8 eligibility boundary cases added after the adversarial checks), and the reference outputs for every model version form the test vectors. The JavaScript scorer is a line-by-line port that must reproduce every vector, including the order of findings, and a test fails if the site’s copies of the model files drift from the released versions. Each scored result lists the contributing features and the supporting chart records.

3.4 Components

Patient demonstration. An installable web application administers the 2024 RECOVER-Adult symptom list [12], with item wording from the RECOVER questionnaire [13], the DePaul Symptom Questionnaire post-exertional malaise items [14], the Post-COVID-19 Functional Status scale [15] (for people who report possible or prior COVID-19), and 5 custom change questions. The app shows the RECOVER total and its research category only when every symptom is answered and otherwise shows the points so far as incomplete. Optional condition self-checks, written for the app from published criteria, cover Long COVID symptom duration (the 2024 National Academies definition [16]), the ME/CFS core symptoms [17], the fibromyalgia Widespread Pain Index and Symptom Severity Scale [18], and possible Lyme disease exposure. Each reports whether the answers match the pattern the criteria describe and is a prompt for the clinician, not a diagnosis. Two further optional parts give the published Patient Health Questionnaire-9 (PHQ-9) [19] and Generalized Anxiety Disorder-7 (GAD-7) [20] totals and severity bands once every item is answered. Any answer other than “not at all” to the PHQ-9 item on thoughts of death or self-harm shows the 988 Suicide and Crisis Lifeline at once, also offline, and puts a notice at the top of the patient’s answers in the clinician view. The app runs in English or Spanish, storing the same answers and computing the same scores in either language. The Spanish version uses the RECOVER study’s Spanish survey wording, the scale authors’ Spanish translation of the functional status scale, and Spanish versions of the DePaul items published by the questionnaire’s developers, and we translated the remaining questions for the app. Answers are stored only in the browser. The current demonstration shows findings from synthetic records in plain language, with dates and without model probabilities or risk multiples. Connection to a person’s own health record remains planned. A printable visit summary lets patients choose what to share, and questionnaire answers can be downloaded as a FHIR R4 QuestionnaireResponse (questionnaire version 1.3.0) with stable item identifiers or handed to the

clinician demonstration within the browser. After initial loading and caching, the questionnaire works offline, but trial searches use the ClinicalTrials.gov API and need a network connection.

Clinician application. A SMART on FHIR application launched with authorization code flow and Proof Key for Code Exchange reads the patient’s conditions, encounters, observations, medication requests, and procedures. A dashboard lists conditions needing review (High or Elevated bands) above a one-line summary of conditions checked and shows the patient’s questionnaire answers and self-check results when present. Each condition has a tab with the band, up to 8 dated findings (more on expansion), each shown with the odds ratio between the estimate with and without it on one scale shared by every patient and model (a full bar is a doubling or halving of the odds) and including findings that lower the estimate, demographic inputs listed separately, criteria to review (the 2024 National Academies Long COVID definition [16], the Institute of Medicine 2015 ME/CFS criteria [17], and the 2016 revised fibromyalgia criteria [18]), and next steps, plus a timeline of chart events. A card whose model could not read the records it needs says “Could not score” and names the missing records, and the summary then notes that a result is missing. The dashboard imports only responses to the app’s own Questionnaire in a supported version and for the open patient, selects the most recent valid one, and attaches a response that names no patient only after the clinician confirms it. When the attached response has a result for a condition’s instrument, the dashboard updates the chart estimate by Bayes’ rule, multiplying the chart odds by that result’s published likelihood ratio, measured against a reference close to the model’s target: the RECOVER index against a U09.9 code (2.83 if positive, 0.46 if negative) [21], questionnaire-based Institute of Medicine criteria against a physician panel’s diagnosis among fatigued adults (1.62, 0.47) [22], and the patient-completed 2016 fibromyalgia criteria against a reported physician diagnosis (8.46, 0.70) [23]. The band, multiple of the average, and next steps then follow the combined estimate, with the chart-only values shown beside it. Instruments without a published likelihood ratio against a comparable reference (the post-exertional malaise screen alone, the Long COVID time criterion, the functional status scale, the Lyme exposure questions, the PHQ-9, and the GAD-7) are shown but not combined. The Lyme disease and lupus tabs show their model’s estimate beside the evidence of a chart rule: for Lyme disease, coded findings consistent with the infection, with historical surveillance incidence for the patient’s county [24] and CDC testing guidance [25], and for lupus, chart evidence mapped to the 2019 classification criteria [26], matching

criterion terms on word boundaries, counting only the longest of overlapping matches, ignoring negated mentions in the same clause, and comparing white-cell and platelet counts in a common unit. A no-login demonstration runs the same code on 4 synthetic personas drawn from the 10 demonstration patients of the test corpus (Figure 2).

Integration service. A CDS Hooks patient-view service declares prefetch templates for the resources the models read and validates each request, including that every supplied or fetched resource belongs to the requested patient. It fetches missing resources with the supplied token only from FHIR servers on a configured allow-list (by default, the public synthetic sandbox) and returns a card for a condition in the Elevated or High band. When records a model needs cannot be read, it returns one information card naming the conditions it could not score instead of an empty card list. It is a synthetic-data prototype: it does not authenticate callers, which CDS Hooks 2.0 specifies through a signed JSON Web Token from the CDS client and which a deployment with real patient data would require [8].

Atlas. The atlas shows one condition at a time (Long COVID, ME/CFS, Lyme disease, lupus, fibromyalgia, migraine, endometriosis, depression, or anxiety): a state map where one was released, how often the condition is coded by age group, sex, and race and ethnicity, and the coded findings more common in people with the condition, all from the companion studies’ aggregate releases. For the self-reported conditions, coding is the share of people who report the diagnosis and have its code, shown beside hypertension and type 2 diabetes in the same group. For Long COVID, coding is the coded share relative to survey prevalence, and the map adds state prevalence from the Household Pulse Survey [27]. The Lyme disease map adds county incidence from CDC surveillance [24]. For ME/CFS, fibromyalgia, and endometriosis, a panel compares Fitbit steps, sedentary and very active minutes, and sleep in the year before the survey between people who report the condition and people who do not, standardized to the same age and sex mix. The atlas shows shares and intervals, never All of Us counts, and its build checks that no hidden group can be recovered from the groups shown. For Long COVID, a scenario calculator uses comparator-model inputs and user-specified assumptions. When the inputs would imply an impossible yield, such as more patients found than reviewed, it raises the dependent input to the nearest feasible value and says so. The atlas does not run the scorer, and the calculator’s outputs are illustrative.

Developer documentation. A documentation page describes each integration path with request

and response examples produced by the running code, runs the scorer in the browser on any test patient and compares the result with the reference output, and lists the model files with their hashes. Its tests check each documented endpoint, parameter, scope, and example against the code.

3.5 Data flow and privacy

Patient answers are stored locally in the browser, and the clinician application processes retrieved records in the browser. The CDS service processes supplied chart data per request and contains no application-level persistence or logging of that data. Hosting-level retention and logging require separate verification. Each page has a content security policy that names the only origins it may contact. Clinician-facing outputs expose supporting information for independent review, but we did not determine the regulatory status of the software’s individual functions.

3.6 Technical evaluation

The evaluation used synthetic records only. It comprised (1) unit and integration tests run with the Node.js test runner and the Python reference tests, (2) cross-implementation comparison of all reference outputs, (3) synthetic CDS Hooks requests, (4) a launch through the public SMART Health IT sandbox, (5) adversarial checks, (6) structural FHIR conformance checks of the test bundles, the example CDS Hooks requests, and the patient app’s QuestionnaireResponses, and (7) a comparison of the condition review that the clinician application shows when launched through the public SMART sandbox with the review the no-login demonstration shows for the same chart. In the adversarial checks, an artificial intelligence (AI) coding agent (Codex, OpenAI), run in a read-only session separate from development, constructed synthetic counterexamples for eligibility, date handling, data availability, and rule extraction, and we reviewed and adjudicated each finding. For check 7, which covered every synthetic patient, we shifted the sandbox charts’ dates so that the scoring date is the upload date. We fixed each failure the adversarial checks found, added its counterexample to the test suite, and reran checks 1 to 3 and 6 on the fixed release. We report each check’s outcome, including failures.

4 Results

4.1 Models

Table 1 lists the deployed models. The deployed Long COVID model contains 69 features, of which 57 have nonzero weights, and had an area under the receiver operating characteristic curve (AUROC) of 0.674 (95% CI, 0.592–0.747) in a random validation split. It is one of the summary-count models compared in the companion analysis, which also exported linear discriminants from the same counts [1]. We deploy naive Bayes because each weight is a per-feature log odds ratio that the application can show. A separately fitted 41-feature model evaluated the development procedure at 18 held-out sites (groups defined by the EHR source of each patient’s index event), with an AUROC of 0.659 (95% CI, 0.576–0.722). We did not externally validate the deployed probability map. The chronic fatigue and fibromyalgia models had AUROCs of 0.735 and 0.726 for self-reporters with no qualifying code in random validation splits, and the Lyme disease and lupus models 0.651 and 0.824. The endometriosis model, scored only for women, reached 0.610. Its discrimination is limited and the ever-diagnosed survey label may not represent current disease, and its clinician card says so. Models for migraine, depression, and anxiety, fitted the same way, reached 0.691, 0.759, and 0.792 for self-reporters with no qualifying code. They are published as model files with reference outputs and are shown in the clinician application and returned by the CDS Hooks service like the other models. We did not assess disease activity.

4.2 Technical evaluation

Table 2 summarizes the checks. The JavaScript suite contained 474 tests: 457 passed, 16 were skipped for two superseded model versions, and 1 was skipped because its build step needs Python in the test runner. Across the 18 synthetic patients (10 demonstration patients and 8 eligibility boundary cases), the nine model files and one placeholder version reproduced all 180 reference outputs with no observed numeric difference, and the current Python reference regenerated the same 180 outputs. Synthetic CDS Hooks requests returned the cards the test vectors predict. Launched through the public SMART sandbox, the application matched the demonstration’s condition review for each of the 18 patients after the comparison found 5 defects. The sandbox uploader shifted a medication code that resembled a date, left a later-dated record inside the scoring window, and

Model (version)	Label and target	Features	Validation: AUROC (95% CI), AUPRC
Long COVID (1.2.0)	First U09.9 code in days 91–455 among adults uncoded at day 90	69 (57 nonzero)	Random 20% split: 0.674 (0.592–0.747), 0.028. Separately fitted 41-feature version at 18 held-out sites: 0.659 (0.576–0.722), 0.018
Chronic fatigue (1.0.0)	Self-reported diagnosis, evaluated among self-reporters with no qualifying code	73	Random 20% split: 0.735 (0.720–0.749), 0.156
Fibromyalgia (1.0.0)	Self-reported diagnosis, evaluated among self-reporters with no qualifying code	80	Random 20% split: 0.726 (0.710–0.743), 0.080
Endometriosis (1.0.0)	Self-reported diagnosis among women, evaluated among self-reporters with no qualifying code	90	Random 20% split: 0.610 (0.595–0.621), 0.111
Lyme disease (1.0.0)	Self-reported diagnosis, evaluated among self-reporters with no qualifying code	86	Random 20% split: 0.651 (0.632–0.673), 0.045
Lupus (1.0.0)	Self-reported diagnosis, evaluated among self-reporters with no qualifying code	79	Random 20% split: 0.824 (0.777–0.863), 0.029

Table 1: Deployed models. We fitted all models as naive Bayes with empirical-Bayes shrinkage outside the enclave from released summary counts and calibrated them from released score bins. Validation is retrospective, in the All of Us Research Program Registered Tier Dataset v9. We have not evaluated any model prospectively or on real FHIR records. Long COVID version 1.2.0 adds the training cohort’s earliest index date (July 2, 2021) and a third excluded SNOMED CT post-COVID concept to version 1.1.0, and we refit its weights and calibration on a corrected data extraction that takes condition codes from both the condition and observation tables. AUPRC, area under the precision-recall curve.

omitted the demonstration’s questionnaire answers. The application rejected the sandbox’s next-page links, and which dated record a finding cited depended on the server’s result order. We fixed all five and added them as tests. One difference remains by design: a COVID-19 diagnosis kept before the model’s earliest index date falls outside the 5-year look-back after the date shift, so one finding lists one date instead of two, with the same weight and score. The structural conformance checks passed for every bundle, example request, and QuestionnaireResponse. We did not run the HL7 FHIR validator.

The adversarial checks, run on the earlier release, found 5 classes of failure. The scorers could score patients outside the intended population (a child, or a patient with a Long COVID code before the COVID-19 index). A valid partial FHIR birth date stopped scoring. Unavailable or failed searches could be treated as absent findings, and the CDS service could return no card

instead of an inability-to-score state. The lupus rule mishandled overlapping terms, negation, and white-cell units. The atlas calculator accepted inputs that implied impossible yields. We fixed all five in the release described here and added each counterexample as a synthetic test. The scorers now enforce the training cohort’s eligibility rules (age 18 years or older, an index on or after July 2, 2021, no Long COVID code by the scoring date, and scoring only from day 90 to day 455) and read partial dates. A failed or forbidden search yields an explicit could-not-score result, and the CDS service returns a card saying the chart could not be scored. The lupus rule uses word-boundary longest matches, negation, and unit normalization, and the calculator enforces joint feasibility. The fixes added 33 tests and the 8 boundary patients. The same review led to request validation, patient-identity checks, and the FHIR server allow-list in the CDS service, to stricter questionnaire import, and to the incomplete RECOVER state. Figure 2 shows the patient and clinician views for synthetic patients.

Check	Outcome
Automated tests (Node.js, Python)	474 JavaScript tests: 457 passed, 17 skipped (16 for superseded model versions). Python reference tests passed
Cross-implementation parity	180 of 180 reference outputs reproduced (18 patients, 10 model versions) with no numeric difference
CDS Hooks (synthetic requests)	Discovery and patient-view requests returned the expected cards. Requests and patient identity validated, and the FHIR fallback limited to an allow-list. Caller authentication not implemented
SMART launch (public sandbox)	All 18 patients’ condition reviews matched the demonstration after we fixed 5 defects (one finding’s date list differs by design). End-to-end launches checked in a browser for 2 patients. Commercial-EHR launches not evaluated
Adversarial checks	5 failure classes found (eligibility, partial dates, unavailable data, lupus rule extraction, calculator feasibility), all fixed and added as tests
FHIR conformance	Structural checks of all bundles, CDS example requests, and QuestionnaireResponses (UUID fullUrls, no null values, R4 required elements). HL7 FHIR validator not run

Table 2: Technical evaluation on synthetic records. The automated tests, parity, CDS Hooks, and conformance outcomes are for the release described here, after the fixes prompted by the adversarial checks. We ran the adversarial checks on the earlier release.

5 Discussion

This prototype shows that models fitted outside a research enclave from released summary counts can be packaged as a versioned contract and delivered through FHIR, SMART on FHIR, and

A.

Explore the patient app

Try it with synthetic patient answers below, or [Try it yourself](#)

Maria Long COVID pat...

Elena Chronic fatigue ...

Sofia Fibromyalgia pat...

Dana Lyme disease co...

Page 1 of 14: About you

HomeNext

About you

Who is answering?

☒ I am answering for myself

☐ I am helping someone else
 Answer the questions for the person you are helping.

When did your symptoms start?

A rough date is fine.

01/20/2024

Have you had COVID-19?

☒ Yes

☐ No

☐ Not sure

About when did you have COVID-19?

A rough date is fine.

01/08/2024

B.

Condition review

Maria Synthetic Female · 44 · as of June 1, 2024

Needs review

Long COVID High 11.6+ average

Top findings: 3+ symptom types coded in days 31-90; Personal history of COVID-19 coded (days 31-90)

Also checked ME/CFS typical · Fibromyalgia typical · Lyme typical · Lupus typical

From the patient

Completed synthetic responses. Answered May 28, 2024. [Remove](#)

[Review responses](#)

Long COVID • ME/CFS Fibromyalgia Lyme Lupus Timeline

High: chart and answers well above the average

Chance of a first Long COVID code (U09.9) 91 to 455 days after COVID-19, compared with the average post-COVID patient at day 90: 11.6 times the average (estimated 11.0%) with the patient's answers; chart alone 4.4 times (4.2%).

Why

Patient's answers: RECOVER 2024 index, 11 or more points May 28, 2024 ▲ +2.83

3+ symptom types coded in days 31-90 Feb 20, 2024 Mar 5, 2024 Mar 25, 2024 ▲ +1.51

Personal history of COVID-19 coded (days 31-90) Mar 12, 2024 ▲ +1.36

4+ outpatient visits in days 31-90 Feb 20, 2024 Mar 5, 2024 Mar 12, 2024 and 1 more ▲ +1.31

Shortness of breath (coded by day 90) Mar 25, 2024 ▲ +1.23

Fatigue or malaise (coded by day 90) Feb 20, 2024 ▲ +1.22

Chest CT Mar 27, 2024 ▲ +1.21

C-reactive protein test Feb 20, 2024 ▲ +1.20

Palpitations, tachycardia, POTS, or orthostatic hypotension (coded by day 90) Feb 20, 2024 ▲ +1.19

7 more findings

Context, not findings: Female, Age 40-49. Age and sex are part of the model but are not chart findings.

From the patient

- Has had COVID-19: Yes (January 8, 2024)
- RECOVER 2024 index: 16 points (11 or more points)
- Post-COVID-19 Functional Status: moderate functional limitations
- DePauli post-exertional malaise screen: meets the screening rule
- Long COVID call-back: 1 visit 3 months or longer (NACSEM 2024 time criterion)

Criteria to check

- Symptoms lasting 3 months or longer after infection
- National Academies of Sciences, Engineering, and Medicine. A Long COVID Definition. 2024.

Figure 2: Patient and clinician views for synthetic patients. **A.** Patient demonstration with the demo personas and the first page of the symptom check, holding one persona’s demonstration answers. **B.** Clinician dashboard for a synthetic patient with a High Long COVID pattern, with the patient’s demonstration answers attached: conditions needing review, the patient’s answers, and the Long COVID tab, where the answers and each dated finding show how much they change the estimate. We cropped both screens to the application area.

CDS Hooks, with one scorer reproducing reference outputs exactly across implementations. Our contribution is the route from an enclave to clinical interfaces, not the models’ accuracy, which is modest and documented in the companion papers.

Separating the model contract from the applications had practical benefits. The test vectors let us check the JavaScript scorer against the reference without any patient data, and adding a model or condition changed only the model file and its vectors. Adding a condition still requires a feature registry, a compliant aggregate extraction, model fitting and calibration, implementation checks, and condition-specific validation. Compared with deployment frameworks that run custom models inside one health system’s infrastructure [4], the contract here is static and portable, which suits enclave-derived models but restricts them to binary features with sufficient counts.

The adversarial checks showed that passing our own tests did not guarantee correct behavior at the edges. Eligibility rules, partial dates, and unavailable data are exactly where a deployed

tool meets real records, and the original test vectors did not cover them. Each counterexample is now a test, so a recurrence of these failures would fail the suite, but the suite covers only the cases someone thought to construct. A review that constructed counterexamples was the most informative part of the evaluation, and we recommend such a review for similar systems. Here, an AI coding agent working apart from the development sessions constructed them, and we checked each finding.

5.1 Limitations

This evaluation has several limitations. First, it used synthetic records only. We did not evaluate real-record interoperability, usability, or prospective clinical performance, and access to commercial EHR sandboxes was pending. Second, the models predict documentation or a self-reported diagnosis rather than independently established disease, and their scores may reflect care access, utilization, coding practices, and incomplete records. Third, we fitted the calibration on training data and have not evaluated it in deployment populations, and the risk bands are heuristic. The update with the patient’s answers treats them as independent of the chart findings given the condition, as the naive Bayes models treat their own features. Where the chart codes the same symptoms the answers count, the combined estimate overstates. None of the likelihood ratios was measured in the application’s population, and because the app’s RECOVER total can only exceed the research index, the published Long COVID ratio likely overstates the app’s. Fourth, we have not validated the equivalence of the FHIR features to the research data model. Participant-provided FHIR data in All of Us hold far fewer condition and procedure records than the research data [28], and the application reads FHIR administrative gender rather than sex at birth. Fifth, our own tests checked FHIR conformance structurally, and we did not run the HL7 FHIR validator. Sixth, the CDS service does not authenticate callers, and records hidden by security labels are treated as readable rather than unavailable. Seventh, the lupus rule detects negation with a short list of cues, and no rheumatologist has reviewed its outputs. Eighth, the RECOVER items use the published questionnaire’s symptom wording [13], but the question stem is ours. The app does not administer the severity measures that the index requires for fatigue, brain fog, and chest pain, so its total can exceed the research index by up to 5 points. The combined questionnaire and chart triage rule and the condition self-checks are exploratory. Finally, we restricted the Post-COVID-19

Functional Status scale to participants who report possible or prior COVID-19, and the custom change questions use symptom onset when COVID-19 does not apply. We have not validated the resulting questionnaire workflow for chronic fatigue case finding.

6 Conclusion

This prototype combines models fitted from released summary counts with a shared scoring implementation, patient and clinician interfaces, a CDS Hooks service, and an aggregate-data atlas. Synthetic testing established reproducible scoring across the reference and JavaScript implementations and identified 5 failure classes at the boundary with real records, which we fixed and covered with tests. Real-record interoperability, usability, and prospective clinical performance remain to be evaluated.

Clinical Relevance Statement

This prototype shows how models fitted from released summary counts can be delivered through patient and clinician interfaces. Its testing establishes technical consistency on synthetic records, not clinical effectiveness. Real-record interoperability, usability, and prospective clinical evaluation are needed before routine use.

Multiple-Choice Questions

1. Why did the authors fit the deployed models outside the research enclave from summary counts?
(a) Summary counts give more accurate models. (b) The enclave does not allow participant-level data or models trained on them to leave. (c) FHIR requires naive Bayes models. (d) Summary counts remove the need for calibration. *Correct answer: (b).* The All of Us Researcher Workbench allows only aggregate results that meet its dissemination policy to leave. A model fitted outside from released counts can be shared.
2. What did the cross-implementation test vectors establish? (a) That the models are clinically valid. (b) That the JavaScript scorer reproduces the reference outputs on synthetic patients. (c) That the system works in commercial EHRs. (d) That the risk bands are calibrated. *Correct*

answer: (b). The vectors check software consistency. The study did not evaluate clinical validity, EHR interoperability, or calibration in deployment.

Funding

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

Conflict of Interest

None declared. [verify](#)

Protection of Human and Animal Subjects

The underlying analyses used de-identified data from the All of Us Research Program Registered Tier Dataset v9 (curated data repository R2025Q4R6, data cutoff January 1, 2025). All of Us participants provide informed consent at enrollment under a protocol overseen by the program’s institutional review board. The program’s data passport model allows registered researchers to analyze de-identified Registered Tier data without project-specific review [verify](#). The system evaluation used synthetic patients only and involved no human subjects.

Declaration of Generative AI Use

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 software in a read-only session separate from development, which produced the adversarial checks described in Methods. 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. A live demonstration with synthetic patients is at <https://undiagnosednomore.org>, with the model files, specification, test vectors, and example CDS Hooks requests on its developer pages. We will release the source code in a public repository (URL and license at posting).

References

- [1] Amir Asiaee and Kaveh Aryan. Long COVID undercoding and early prediction of its documentation: survey comparison and an exportable model in the All of Us Research Program, 2026. Companion manuscript; medRxiv preprint forthcoming.
- [2] 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.

- [3] All of Us Research Program. Data and statistics dissemination policy. <https://support.researchallofus.org/hc/en-us/articles/22346276580372-Data-and-Statistics-Dissemination-Policy>, 2026. Accessed October 8, 2026.
- [4] Conor K. Corbin, Rob Maclay, Aakash Acharya, Sreedevi Mony, Soumya Punathanam, Rahul Thapa, Nimesh 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.
- [5] HL7 International. FHIR Release 4. <https://hl7.org/fhir/R4/>, 2019. Version 4.0.1, October 30, 2019; accessed October 8, 2026.
- [6] 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.
- [7] HL7 International. SMART App Launch implementation guide. <https://hl7.org/fhir/smart-app-launch/>, 2024. Version 2.2.0 (STU 2.2), April 30, 2024; accessed October 8, 2026.
- [8] HL7 International. CDS Hooks 2.0. <https://cds-hooks.hl7.org/2.0/>, 2022. Version 2.0 (STU 2), August 23, 2022; accessed October 8, 2026.
- [9] Jan Talmon, Elske Ammenwerth, Jytte Brender, Nicolette de Keizer, Pirkko Nykänen, and Michael Rigby. STARE-HI—Statement on reporting of evaluation studies in Health Informatics. *International Journal of Medical Informatics*, 78(1):1–9, 2009. doi: 10.1016/j.ijmedinf.2008.09.002.
- [10] 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/NEJMSr1809937.
- [11] Amir Asiaee and Kaveh Aryan. Self-reported diagnoses as training labels for finding uncoded chronic fatigue, fibromyalgia, and endometriosis in the All of Us Research Program, 2026. Companion manuscript; medRxiv preprint forthcoming.

- [12] Linda N. Geng, Kristine M. Erlandson, Mady Hornig, et al. 2024 update of the RECOVER-Adult Long COVID Research Index. *JAMA*, 333(8):694–700, 2025. doi: 10.1001/jama.2024.24184.
- [13] Tanayott Thaweethai, Sarah E. Jolley, Elizabeth W. Karlson, et al. Development of a definition of postacute sequelae of SARS-CoV-2 infection. *JAMA*, 329(22):1934–1946, 2023. doi: 10.1001/jama.2023.8823.
- [14] Joseph Cotler, Carly Holtzman, Catherine Dudun, and Leonard A. Jason. A brief questionnaire to assess post-exertional malaise. *Diagnostics*, 8(3):66, 2018. doi: 10.3390/diagnostics8030066.
- [15] Frederikus A. Klok, Gudula J. A. M. Boon, Stefano Barco, Matthias Endres, J. J. Miranda Geelhoed, Samuel Knauss, Spencer A. Rezek, Martijn A. Spruit, Jörg Vehreschild, and Bob Siegerink. The Post-COVID-19 Functional Status scale: a tool to measure functional status over time after COVID-19. *European Respiratory Journal*, 56(1):2001494, 2020. doi: 10.1183/13993003.01494-2020.
- [16] National Academies of Sciences, Engineering, and Medicine. *A Long COVID Definition: A Chronic, Systemic Disease State with Profound Consequences*. National Academies Press, Washington, DC, 2024. doi: 10.17226/27768.
- [17] Institute of Medicine. *Beyond Myalgic Encephalomyelitis/Chronic Fatigue Syndrome: Redefining an Illness*. National Academies Press, Washington, DC, March 2015. ISBN 9780309316897. doi: 10.17226/19012.
- [18] Frederick Wolfe, Daniel J. Clauw, Mary-Ann Fitzcharles, Don L. Goldenberg, Winfried Häuser, Robert L. Katz, Philip J. Mease, Anthony S. Russell, Irwin Jon Russell, and Brian Walitt. 2016 revisions to the 2010/2011 fibromyalgia diagnostic criteria. *Seminars in Arthritis and Rheumatism*, 46(3):319–329, 2016. doi: 10.1016/j.semarthrit.2016.08.012.
- [19] Kurt Kroenke, Robert L. Spitzer, and Janet B. W. Williams. The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613, 2001. doi: 10.1046/j.1525-1497.2001.016009606.x.

- [20] Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the GAD-7. *Archives of Internal Medicine*, 166(10):1092–1097, 2006. doi: 10.1001/archinte.166.10.1092.
- [21] 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.
- [22] Leonard A. Jason, Madison Sunnquist, Bobby Kot, and Abigail Brown. Unintended consequences of not specifying exclusionary illnesses for systemic exertion intolerance disease. *Diagnostics*, 5(2):272–286, 2015. doi: 10.3390/diagnostics5020272.
- [23] Sachin Srinivasan, Eamon Maloney, Brynn Wright, Michael Kennedy, K. James Kallail, Johannes J. Rasker, Winfried Häuser, and Frederick Wolfe. The problematic nature of fibromyalgia diagnosis in the community. *ACR Open Rheumatology*, 1(1):43–51, 2019. doi: 10.1002/acr2.1006.
- [24] Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases, Division of Vector-Borne Diseases. Lyme disease public use aggregated data with geography, 2008–2021. data.cdc.gov dataset qtbi-xd4i, <https://data.cdc.gov/d/qtbi-xd4i>, 2025. Accessed September 28, 2026.
- [25] Centers for Disease Control and Prevention. Clinical testing and diagnosis for Lyme disease. <https://www.cdc.gov/lyme/hcp/diagnosis-testing/index.html>, 2024. Page last reviewed May 15, 2024; accessed October 4, 2026.
- [26] Martin Aringer, Karen Costenbader, David Daikh, et al. 2019 European League Against Rheumatism/American College of Rheumatology classification criteria for systemic lupus erythematosus. *Arthritis & Rheumatology*, 71(9):1400–1412, 2019. doi: 10.1002/art.40930.
- [27] National Center for Health Statistics. Post-COVID conditions. <https://data.cdc.gov/d/gsea-w83j>, 2024. Household Pulse Survey estimates, June 1, 2022, to September 16, 2024; accessed September 28, 2026.

- [28] 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.