

PHASE-ED: Recorded Emergency Department Stay and Temporal Prediction Across Pandemic Phases at Jordanian Public Hospitals

Samer Abaddi

eng.sameralabbadi@gmail.com

IDentity Research <https://orcid.org/0000-0003-4612-7384>

Research Article

Keywords: Emergency department, Length of stay, COVID-19, Jordan, Electronic health records, Quantile regression, Machine learning, Prediction intervals

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11366553/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

PHASE-ED: Recorded Emergency Department Stay and Temporal Prediction Across Pandemic Phases at Jordanian Public Hospitals

Samer Abaddi

IDentity Research, Amman, Jordan

ORCID: <https://orcid.org/0000-0003-4612-7384>

Correspondence: Samer Abaddi, IDentity Research, Amman, Jordan. Email: eng.sameralabbadi@gmail.com

Abstract

Shorter recorded emergency department stays may reflect changes in care, patient composition or documentation. PHASE-ED characterised recorded stays across pandemic phases, examined recording-period sensitivity and assessed prediction using sparse variables in Jordanian Ministry of Health records. Following a facility and coverage audit of 2017–2025 data, the primary retrospective series comprised 382,018 eligible encounters from 85,356 patients at one facility during 2018–2025; seven other facility codes contributed separately bounded analyses. Median recorded length of stay was 63.0 minutes before COVID, 60.0 during COVID and 51.0 during recovery. Corresponding 90th percentiles were 317.0, 294.0 and 156.6 minutes, and proportions exceeding six hours were 8.33%, 7.73% and 1.22%. Recovery medians were already lower before an observed timestamp-precision transition on 23 July 2024. Historical and recent benchmarks, quantile regression and boosted trees were developed chronologically and evaluated on 191,497 later encounters. Test mean absolute error was 49.25 minutes for the historical benchmark, 47.66 for regression and 48.02 for boosted trees; the development simplicity rule retained the historical benchmark. Its nominal 80% interval covered 91.5% overall but only 68.7% within the recorded Resuscitation category. Thus, shorter recovery intervals coexisted with modest predictive gains and uneven uncertainty coverage.

Keywords: Emergency department; Length of stay; COVID-19; Jordan; Electronic health records; Quantile regression; Machine learning; Prediction intervals

Introduction

The time spent in an emergency department (ED) connects a patient's immediate need for care with the hospital's ability to provide it. Length of stay (LOS) reflects assessment and treatment alongside access to investigations, specialist advice and onward care. The input–throughput–output model places these processes within a wider system of demand and hospital capacity (Asplin et al., 2003). A study involving more than 5.2 million admitted patients found that longer waits for inpatient admission were associated with higher adjusted 30-day mortality beyond 5 hours from ED arrival (Jones et al., 2022). Although observational, this evidence makes time in emergency care a concern extending beyond administrative efficiency. It is a matter of life or death.

COVID-19 disrupted the relationship between attendance, case complexity and available capacity, and subsequent trajectories have differed across settings. Among older US adults, the proportion of ED encounters exceeding 8 hours increased from 12% in 2017 to 20% in 2024, with only partial improvement during 2022–2024 (Haimovich et al., 2025). A recent study from the Middle East and North Africa (MENA) of 69,870 encounters found a modest reduction in mean LOS after lockdown, from 330.5 to 319.2 minutes, alongside changes in age, acuity and admission patterns (Faris et al., 2026). These findings make longer follow-up necessary and caution against treating shorter stays, higher attendance or the lifting of restrictions as interchangeable evidence of recovery.

Interpretation also depends on what the information system measures. Arrival-to-room, room-to-disposition and admission-decision-to-departure intervals represent different operational processes; Canellas et al. (2024) modelled these components separately across 4 EDs. An administrative endpoint cannot establish which process has improved unless its meaning is known. Direct observation has also shown that electronic-record implementation can change the accuracy of operational timestamps (Ward et al., 2013). Consequently, an

apparent temporal improvement may combine changes in care, patient composition, endpoint completeness and recording practice. For longitudinal research, continuity of measurement is part of the scientific question.

Jordan provides an important setting in which to examine these issues. Official statistics record 32 Ministry of Health (MOH) hospitals with 6,059 hospital beds in 2024, situating public ED services within a substantial hospital network (Department of Statistics, 2026). During the initial pandemic response, the allocation of King Abdullah University Hospital to COVID-19 care coincided with altered service availability and ED activity (Al-Qudah et al., 2022). Recently, Rawshdeh et al. (2025) analysed public-hospital electronic records from 2018–2023, describing attendance, hospitalisation and categorical ED-duration patterns. At Jordan University Hospital, Al-Na'seh et al. (2024) examined investigations, consultation, admission and documentation delays during a quality-improvement project conducted in 2022. Mean consultation time decreased from 91 to 65 minutes across its improvement cycles; this was a process-specific measure rather than total ED LOS. These studies establish the relevance of administrative records and locally actionable delays. Their respective national utilisation and short intervention perspectives leave a different question open: how consistently can recorded stays be compared within a public facility over successive pandemic and recovery periods?

Concerning prediction, information must be evaluated in relation to when it could become available. Shbool et al. (2023) demonstrated a Jordanian precedent for LOS classification using 400 selected records from 2019–2020, including medication, investigation and consultation information. That task differs from predicting a continuous interval using a restricted set of arrival-related variables and testing it in later years. More broadly, Farimani et al. (2024) identified 34 ED-LOS prediction studies, comprising 29

development studies and 5 validation studies, with weaknesses in missing-data handling, reproducibility and other design choices.

The gap addressed here lies at the intersection of these literatures. The selected studies in Table 1 establish service disruption, local delays, administrative utilisation and predictive feasibility. They do not jointly establish whether within-facility LOS distributions remain comparable through later recovery years, whether observed changes withstand recording-period checks, and whether predictions retain useful accuracy and uncertainty estimates in subsequent encounters. Overall error alone cannot show whether a prediction interval is sufficiently informative or whether its coverage differs across clinically relevant recorded groups. Addressing these questions requires linking source auditing to description and prediction within the same study.

PHASE-ED uses Demographics and ED Log from Jordanian MOH records to address this problem. Following assessment of the 2017–2025 source coverage, the primary longitudinal analysis concerns one facility during 2018–2025; seven additional facility codes contribute separate supporting analyses within their available windows. The primary objective is to describe recorded arrival-to-departure LOS across pre-COVID, COVID and recovery periods and examine its sensitivity to eligibility rules, observation windows, patient subgroups and recording periods. The secondary objective is to compare historical benchmarks, quantile regression and boosted trees using chronological development, calibration and later testing, with attention to both point error and prediction-interval performance. Predictors comprise recorded age, sex and calendar information; acuity supports retrospective subgroup assessment.

The study contributes at several levels. Conceptually, it treats continuity of measurement as a prerequisite for interpreting change in patient flow and service performance. Empirically, it extends the local evidence with an 8-year primary series and

separately bounded supporting observations. Methodologically, it combines cleaning and recording checks with benchmark-led temporal prediction and subgroup assessment of uncertainty. For service managers, the resulting distributions and sensitivity analyses can identify periods or groups requiring investigation. For information-system governance, the audit identifies definitions and recording practices that require confirmation before apparent improvements or model outputs inform operational decisions.

The Methods section describes the data, cohort construction, period definitions and analytical procedures. The Results section presents the longitudinal patterns, sensitivity analyses and secondary prediction findings. The Discussion section interprets their implications for Jordanian emergency care, examines the measurement and generalisability limitations, and identifies priorities for further validation.

Methods

Study design and data sources

PHASE-ED was a retrospective observational study of recorded emergency department (ED) length of stay (LOS), with a secondary prediction analysis. Demographics and ED Log were frozen in a private BigQuery research snapshot. This snapshot date does not establish the clinical extraction cutoff. The source audit covered 14,550,462 ED records and 2,756,052 demographic rows; 12,982,400 arrivals fell within 2017–2025 across 35 facility codes. Coverage and endpoint completeness were examined by facility and month. Facility 273 during 2018–2025 provided the primary longitudinal series; seven other codes contributed separate analyses within audited windows (Supplementary Table S1). Selection prioritised sustained coverage and endpoint completeness without establishing complete attendance capture. Each supporting window followed the source audit. The study did not assume a balanced ten-hospital panel. Reporting was informed by RECORD and TRIPOD+AI (Benchimol et al., 2015; Collins et al., 2024).

Table 1: Selected studies informing the PHASE ED research gap

Study	Setting and observation period	Design and focus	Relevant finding or approach	Question taken forward in PHASE ED
Al-Qudah et al. (2022)	Jordan KAUH (2018–2020)	Hospital activity review; early COVID allocation policy	Documented reduced ED activity and changes in service availability during the initial response.	How do recorded intervals evolve over a longer recovery period in MOH records?
Shbool et al. (2023)	Jordan Arab Medical Centre (2019–2020)	LOS classification; 400 selected encounters	Used clinical and process information, including investigations and consultations, to classify LOS.	How useful are continuous predictions and intervals when development and testing are separated chronologically?
Al-Na'seh et al. (2024)	Jordan Jordan University Hospital March–May 2022	Quality improvement; 3 cycles	Mean consultation time decreased from 91 to 65 minutes; total LOS was a separate measure.	Do longer-term LOS patterns persist across eligibility, subgroup and recording-period checks?
Canellas et al. (2024)	United States 4 EDs One-year sample	Operational prediction; 135,044 encounters	Separated arrival-to-room, room-to-disposition and admission-decision-to-departure intervals.	What conclusions are supportable when only recorded arrival and departure endpoints are available?
Rawshdeh et al. (2025)	Jordan Public hospitals 2018–November 2023	Electronic-record utilisation and hospitalisation analysis	Described attendance, demographic characteristics and categorical ED-duration patterns.	Can within-facility phase comparisons and later prediction be interpreted alongside recording continuity through 2025?
Faris et al. (2026)	Lebanon AUBMC (2020–2022)	Lockdown comparison; 69,870 encounters	Mean LOS declined modestly while age, acuity and admission patterns changed.	How do a longer Jordanian series and subgroup checks contextualise apparent recovery?
PHASE ED Present study	Jordan Primary facility 2018–2025; 7 supporting codes	Recorded-interval description and secondary chronological prediction	Links source auditing, phase comparisons, sensitivity analyses and prediction-interval evaluation.	Provides context-specific evidence; clinical timestamp meanings and generalisability remain unresolved.

ED = emergency department; LOS = length of stay; KAUH = King Abdullah University Hospital; AUBMC = American University of Beirut Medical Center. PHASE ED comprises a primary facility series and separate supporting windows.

Eligibility and outcome definition

Encounter keys and demographic patient keys were verified as unique before linkage. All scoped encounters matched a demographic record. Within the selected facility window, exclusions followed a fixed sequence: invalid identity, missing or unparseable departure, negative duration, exact zero, and otherwise-positive encounters coded ‘Error’. Repeat visits and unknown disposition were retained. Missing outcomes were neither imputed nor assumed censored; negative intervals were not reversed or shifted. Positive durations had no upper cap, and eligible departures after 2025 were retained. Figure 1 presents the resulting cohort flow.

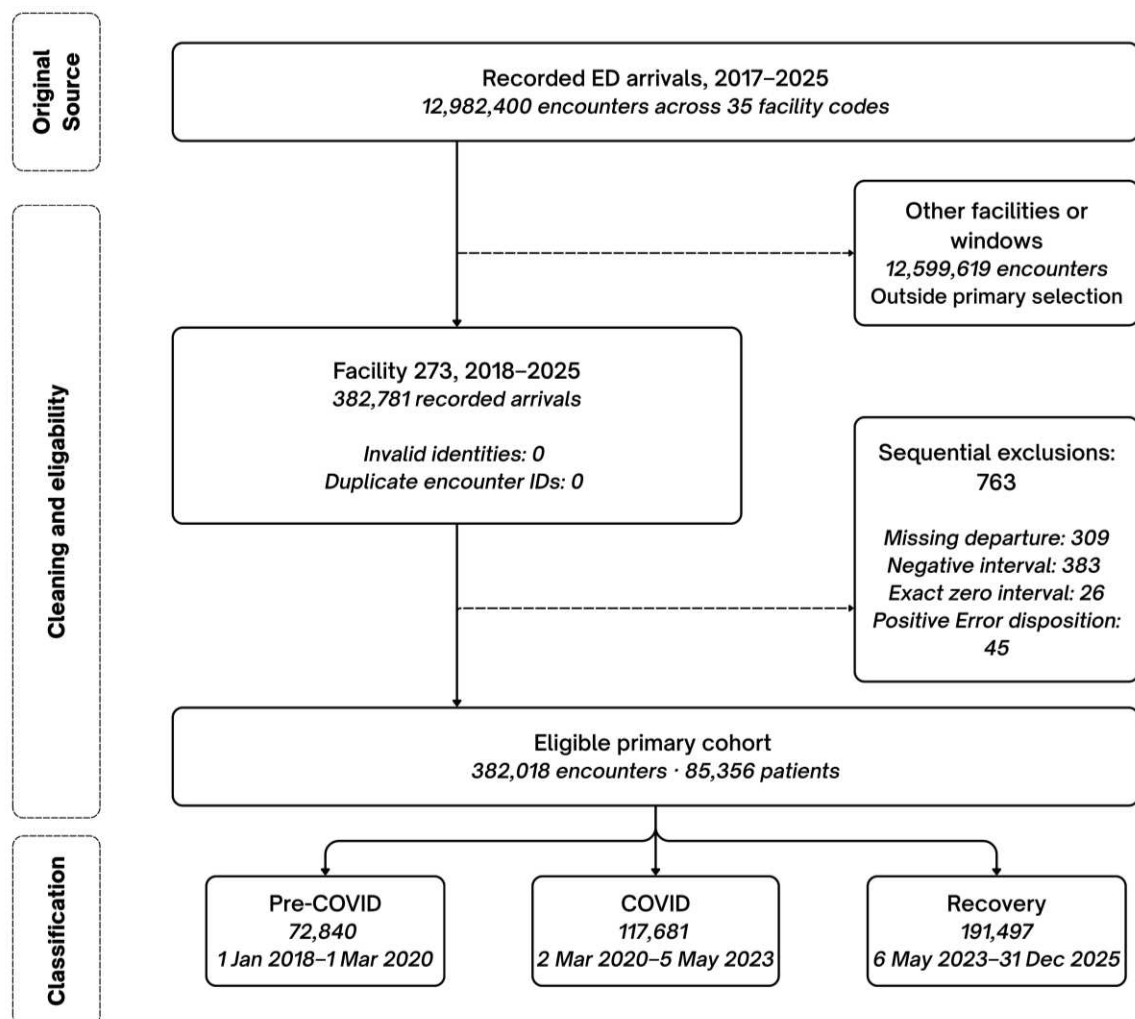


Figure 1: Selection of the primary cohort from audited emergency department records

Recorded LOS in minutes was calculated without rounding using Equation (1), where D and A denote departure and arrival timestamps and $\Delta\mu s$ is their difference in microseconds. Timestamps were treated as stored civil-clock values without inferred timezone conversion. Their clinical event meanings remain unconfirmed. A precision transition beginning on 23 July 2024 defined a recording-period check, without attributing it to a verified system change.

$$y_i = \frac{\Delta_{\mu s}(D_i, A_i)}{60,000,000} \text{ minutes} \quad (1)$$

Periods and analytical variables

In the primary cohort, arrival dates assigned encounters to pre-COVID (1 January 2018–1 March 2020), COVID (2 March 2020–5 May 2023) or recovery (6 May 2023–31 December 2025). These boundaries reference Jordan’s first reported cases and the end of the WHO public health emergency designation (World Health Organization, 2021, 2023). Recovery is an analytical label. Figure 2 aligns these periods with model development and evaluation.

Table 2 distinguishes patient, encounter, facility and calendar variables. Descriptive age used a plausible birth-date fallback, and marked discrepancies exceeding one year between two plausible ages as missing. Modelling instead used plausible encounter-recorded age, a training-fold median for missing values and an explicit missingness indicator. Sex and calendar categories used fixed encodings. Acuity’s scale and initial-assessment timing were unconfirmed; it supported retrospective subgroup checks only. Governorate described the recorded facility location (Hospital names were encrypted from the source). Continuous age was grouped only for descriptive and subgroup summaries, preserving its numerical information during prediction.

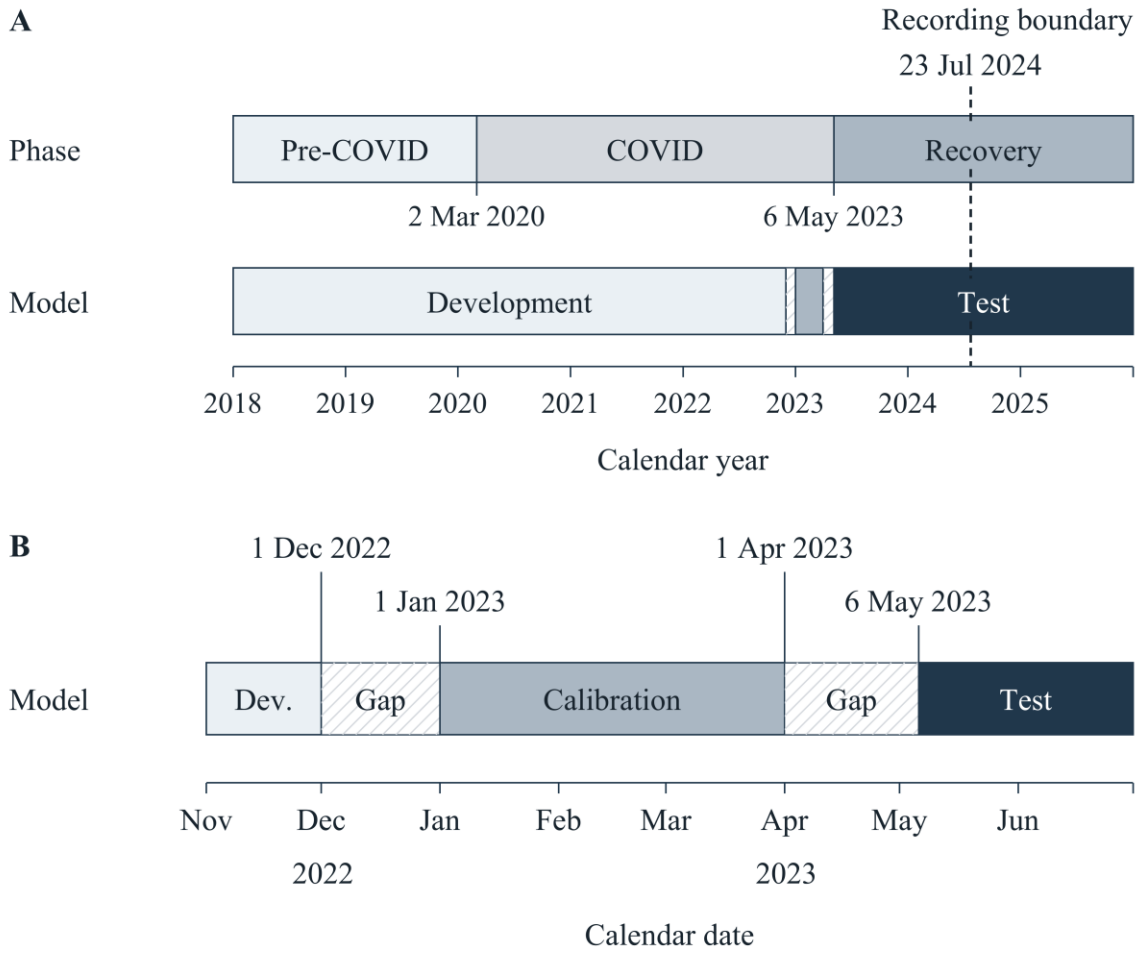


Figure 2: Alignment of pandemic phases with chronological model development and evaluation

Note. A: full primary observation period. B: expanded November 2022–June 2023 view. Dev. = development. Intervals are assigned by recorded arrival date. The expanded view makes the December 2022 and 1 April–5 May 2023 follow-up gaps visible; gap encounters remain eligible for descriptive analyses but do not enter model development, calibration or primary testing.

Table 2: Analytical roles and treatment of the available fields

Domain and field	Type	Treatment	Analytical role
Outcome Arrival and departure	Datetime; continuous LOS	Parse trimmed values; preserve null, blank and parse-failure flags. Retain unrounded positive minutes without a primary upper cap.	Defines eligibility and outcome; departure and derived outcome flags are excluded from predictors.
Patient Age	Continuous; descriptive bands	Accept 0–120 years. Descriptive bands: <18, 18–39, 40–64, ≥65, Unknown. Descriptive reconciliation can use birth date; modelling uses recorded age only, training-median imputation and missingness indicator.	Descriptive and predictive; cubic spline in M1 and raw imputed age in M2.
Patient	Categorical	Male, Female, Unknown. Null, blank and	Descriptive and predictive; fixed

Sex		unmapped values map to Unknown. Current demographic snapshot; historical availability unverified.	one-hot reference coding.
Encounter Acuity	Categorical	Resuscitation, Emergency, Urgent, Less Urgent, Non Urgent, Unknown. Preserve source labels; codes are not equal-interval severity scores.	Retrospective description and model evaluation only; excluded as a predictor.
Encounter Disposition	Categorical	Exclude otherwise-eligible '(10) Error' records in the primary cohort. Retain unknown disposition as a distinct category.	Eligibility, description and sensitivity; excluded as a predictor.
Facility Facility code and governorate	Categorical	Retain anonymous codes and separate audited windows. Governorate is a source-labelled facility attribute with unverified administrative mapping.	Scope and contextual description; neither is a predictor in the single-facility model.
Calendar Arrival time	Categorical	Hour 0–23, weekday 1–7, month 1–12; fixed reference coding. Phase, year and recording period remain separate evaluation fields.	Hour, weekday and month are predictors. Phase and recording period support stratification; year is excluded.
Linkage and other fields	Identifiers, dates and text	Patient and encounter keys support linkage and repeat-attendance checks. No complaint embeddings, bed-capacity interpretation, history or pressure features were constructed.	Identifiers, birth/death dates, complaint text and bed identifiers are excluded from predictors. Orders and Problems are outside the study inputs.

ED = emergency department; LOS = length of stay. Missing predictors did not remove an otherwise eligible encounter.

Descriptive comparisons

LOS distributions were summarised using medians, interquartile ranges, 90th percentiles, means, standard deviations and proportions exceeding 6 hours; 4- and 12-hour summaries were also calculated. Percentiles used linear interpolation on encounter-level observations; phase medians were not averages of monthly medians. Monthly activity and recording-quality indicators accompanied phase comparisons. Contrasts were unadjusted and noncausal, with no hypothesis tests or confidence intervals. Age, sex, acuity and recording-period analyses were exploratory and specified after the primary audit. Sensitivities examined zero and Error-coded intervals, long-duration exclusions, chronology flags, activity and completeness screens, alternative periods and supporting-facility windows (Supplementary Table S2).

Prediction development and selection

Development comprised 1 January 2018–30 November 2022, calibration 1 January–31 March 2023, and testing 6 May 2023–31 December 2025. These sets contained 168,836, 12,274 and 191,497 encounters, respectively. Intervening follow-up gaps separated the stages; their 9,411 eligible encounters were not used for fitting, calibration or testing. Development and calibration outcomes required departures before 1 January and 6 May 2023, respectively. Patients could recur across stages. Three expanding development folds validated on January–November 2020, 2021 and 2022. Chronological benchmarking also has a cross-domain precedent in Abaddi’s (2026) load-forecasting study; PHASE-ED predicted individual intervals rather than aggregate demand.

Historical empirical quantiles (M0), final-90-day training quantiles (M0R), quantile regression with a cubic age spline (M1), and LightGBM quantile trees (M2) were compared. Predictors were age, age missingness, sex, hour, weekday and month. A separately tuned calendar-only tree model was diagnostic. All learned preprocessing used training data alone. Four regression penalties and 12 tree configurations were evaluated per fold; pooled validation mean absolute error (MAE) selected settings. The simplest family within 1% of the smallest MAE was preferred in the fixed order M0, M0R, M1, M2. Selected median-model settings were retained for the 0.10 and 0.90 quantiles. M1 used five age-spline knots, degree 3 and L1 penalty candidates 0, 0.01, 0.1 and 1. M2 fixed 500 trees and learning rate 0.05, tuning leaves (7, 15, 31), minimum child sizes (100, 500) and L2 penalties (0, 1). Supplementary Table S3 outlines the model specifications, complete estimator settings and development selection.

Interval calibration and evaluation

Negative quantile predictions were clipped to zero. The minimum and maximum of the three clipped quantiles formed a noncrossing envelope, while the clipped median

remained the point prediction. Following the quantile-calibration principle of Romano et al. (2019), separate calibration used expansion-only scores and a nominal 80% interval (Equation (2)). Here, L and U denote envelope limits, y the observed interval, and d the 9,820th ordered score among 12,274 calibration encounters. Calibration could widen but not narrow intervals. Repeated attendance and temporal shift precluded an exchangeability-based coverage guarantee.

$$\begin{aligned}
s_i &= \max\{L_i - y_i, y_i - U_i, 0\} \\
d &= s_{([0.80(n_{\text{cal}}+1)])} \\
I_{0.80}(x) &= [\max\{0, L(x) - d\}, U(x) + d]
\end{aligned} \tag{2}$$

Evaluation measured MAE, median absolute error, root mean squared error, pinball loss, interval coverage, width and interval score. Lower and upper miss rates, quantile crossings and predicted-versus-observed summaries were reported separately. No resampling-based uncertainty estimates were computed. Quarterly, recording-period and patient-subgroup results assessed variation; these were descriptive checks rather than formal fairness tests. Additional analyses used patient-disjoint tests, a stricter earlier-arrival exclusion, omission of the final test quarter, Error-inclusive testing, and regression/tree refits excluding development durations above 72 hours. Historical test-outcome distributions had been inspected before protocol specification, although test prediction performance was withheld during selection. The experiment was neither preregistered nor prospective.

Computing and research governance

Models ran privately in Google Cloud Shell using Python 3.12.3, scikit-learn 1.9.0 and LightGBM 4.7.0, four CPU threads and seed 20261012. The host reported approximately 15.6 GiB memory. Synthetic resource tests fixed the regression solver as highs-ipm before study fitting. Supplementary Table S4 records software, hardware limits and execution provenance. Patient rows, fitted models and individual predictions remained in Google

Cloud; only aggregates were exported. Versioned code and execution receipts retained source hashes. The secondary data source is the Jordanian Ministry of Health, via the Hakeem electronic health solutions (EHS). EHS provides data to researchers in an encrypted format, masking sensitive patient information—such as names, hospital details, and any medical specifics (EHS, 2026). The data is handled with strict confidentiality and stored in a secure Google Cloud (monthly subscription for enhanced space) accessible only on unrecognized devices through at least three of the following: (i) a password, (ii) fingerprint scan, (iii) facial recognition, (iv) 6-digit code on Authenticator, (v) OTP on verified phone number(s) only, alongside security alerts for any suspicious activity.

Results

Cohort and recorded characteristics

The primary series comprised 382,018 eligible encounters from 85,356 patients at facility 273 during 2018–2025, representing 99.80% of 382,781 recorded arrivals. Sequential exclusions were 309 missing departures, 383 negative intervals, 26 exact-zero intervals and 45 otherwise-positive Error-coded records. No primary identity failed or required duplicate removal. All eligible encounters linked to Demographics, and repeat visits remained in the cohort. Records occurred in all 96 months, on 2,910 of 2,922 calendar days; days without records did not establish zero attendance. Median age was 27 years in eligible encounters and 28 among the 763 excluded encounters. The three phases contributed 72,840, 117,681 and 191,497 encounters, respectively (Table 3).

Table 3: Recorded characteristics of eligible primary encounters

Characteristic	Overall	Pre-COVID	COVID	Recovery
Eligible encounters	382,018	72,840	117,681	191,497
Distinct patients	85,356	34,542	41,721	55,629
Age, median (Q1–Q3), years	27.0 (12.0–42.0)	26.0 (12.0–42.0)	28.0 (14.0–43.0)	27.0 (12.0–42.0)
Age band: <18	125,356 (32.81%)	23,650 (32.47%)	36,109 (30.68%)	65,597 (34.25%)
Age band: 18-39	146,776 (38.42%)	28,893 (39.67%)	46,421 (39.45%)	71,462 (37.32%)

Age band: 40-64	86,726 (22.70%)	15,449 (21.21%)	27,560 (23.42%)	43,717 (22.83%)
Age band: 65+	23,156 (6.06%)	4,844 (6.65%)	7,591 (6.45%)	10,721 (5.60%)
Age band: Unknown	4 (<0.01%)	4 (<0.01%)	0 (0.00%)	0 (0.00%)
Sex: Female	159,427 (41.73%)	30,227 (41.50%)	49,973 (42.46%)	79,227 (41.37%)
Sex: Male	222,591 (58.27%)	42,613 (58.50%)	67,708 (57.54%)	112,270 (58.63%)
Acuity: (01) Resuscitation	3,326 (0.87%)	499 (0.69%)	1,792 (1.52%)	1,035 (0.54%)
Acuity: (02) Emergency	19,685 (5.15%)	2,808 (3.86%)	11,261 (9.57%)	5,616 (2.93%)
Acuity: (03) Urgent	43,301 (11.33%)	11,012 (15.12%)	13,744 (11.68%)	18,545 (9.68%)
Acuity: (04) Less Urgent	228,027 (59.69%)	33,546 (46.05%)	63,364 (53.84%)	131,117 (68.47%)
Acuity: (05) Non Urgent	87,675 (22.95%)	24,971 (34.28%)	27,520 (23.39%)	35,184 (18.37%)
Acuity: Unknown	4 (<0.01%)	4 (<0.01%)	0 (0.00%)	0 (0.00%)
Disposition: coded unknown	82,173 (21.51%)	19,069 (26.18%)	59,040 (50.17%)	4,064 (2.12%)

Note: Unknown sex is zero throughout. Values below 0.01% are displayed as <0.01%, not 0.00%.

Median age was 26 years before COVID, 28 during COVID and 27 during recovery; the recorded female proportion was 41.50%, 42.46% and 41.37%. Four out-of-range recorded ages and four acuity values were unknown, all before COVID; no sex values were unknown. Zero age was retained in 15,439 encounters. Recorded composition nevertheless changed. Less Urgent increased from 46.05% before COVID to 68.47% during recovery, while Emergency rose from 3.86% to 9.57% during COVID before declining to 2.93%. Unknown disposition was 26.18%, 50.17% and 2.12% across the phases.

Recorded stay across phases

Median recorded LOS was 63.0 minutes before COVID, 60.0 during COVID and 51.0 during recovery (Table 4). The corresponding interquartile ranges were 21.0–147.0, 25.0–132.0 and 24.0–93.1 minutes. Recovery was 12.03 minutes below pre-COVID and 9.03 below COVID when differences were calculated from unrounded medians. The 90th percentile fell from 317.0 to 294.0 and 156.6 minutes; proportions exceeding six hours were 8.33%, 7.73% and 1.22%. Means were 143.66, 120.13 and 72.69 minutes, reflecting the longer right tail in the earlier records.

Table 4: Recorded LOS by study phase and recovery recording period

Period	Encounters	Patients	Median (Q1–Q3), min	P90, min	Mean (SD), min	>6 h, n (%)
Overall	382,018	85,356	55.0 (24.0–111.0)	215.0	100.8 (253.8)	17,505 (4.58%)
Pre-COVID	72,840	34,542	63.0 (21.0–147.0)	317.0	143.7 (514.6)	6,070 (8.33%)
COVID	117,681	41,721	60.0 (25.0–132.0)	294.0	120.1 (179.9)	9,092 (7.73%)
Recovery	191,497	55,629	51.0 (24.0–93.1)	156.6	72.7 (78.5)	2,343 (1.22%)
Recovery: before 23 Jul 2024	72,589	30,672	50.0 (22.0–94.0)	159.0	73.4 (84.4)	1,065 (1.47%)
Recovery: from 23 Jul 2024	118,908	42,317	51.2 (25.4–92.6)	154.9	72.2 (74.6)	1,278 (1.07%)

The phase summaries obscured substantial monthly variation (Figure 3). Within the COVID period, monthly medians ranged from 85 minutes in September 2021 to 43 in December 2022. The lower recovery median therefore followed changes already visible before the end of the COVID phase. These summaries concern recorded durations and do not identify when, or why, clinical processes changed.

Recording periods and subgroup patterns

The recovery median was 50.0 minutes among 72,589 encounters before 23 July 2024 and 51.2 among 118,908 from that date. Six-hour exceedance was 1.47% and 1.07%, respectively (Table 5). The lower recovery median thus preceded the observed timestamp-precision boundary. Non-minute-aligned departures accounted for 116,288 of 119,300 parsed later departures (97.48%), compared with none of 72,597 earlier recovery departures. All 383 primary negative intervals occurred in the later period and were within one hour of zero; 11 were within one minute. The audit established this recording pattern but did not establish its cause (Supplementary Table S1).

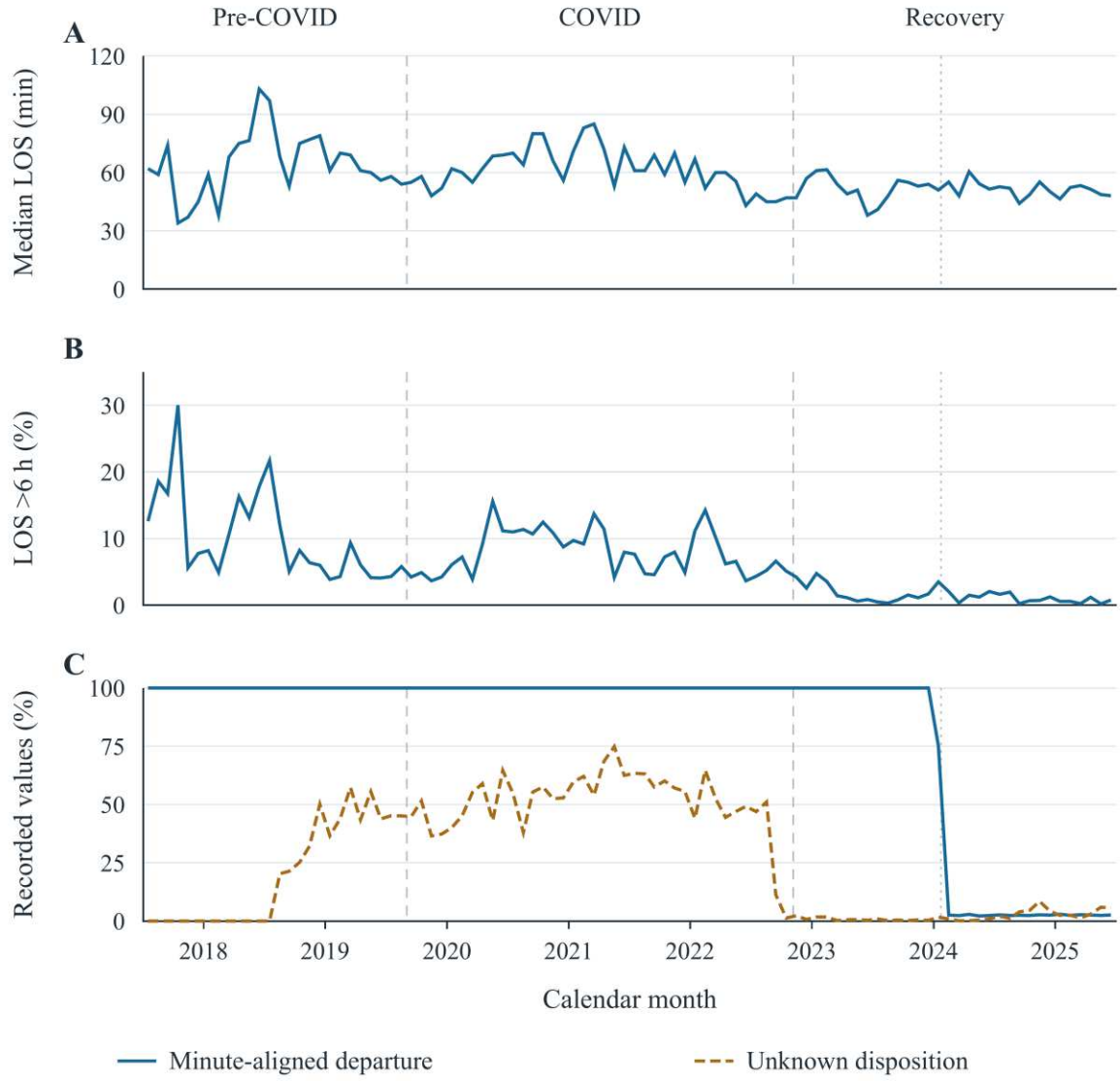


Figure 3: Monthly recorded stay and changing documentation at the primary facility

A: median LOS. B: encounters with LOS >6 hours. C: minute-aligned departures and unknown disposition. Percentages use eligible monthly encounters. Dashed lines mark phase boundaries; the dotted line marks 23 July 2024.

Table 5: Development selection and chronological test performance

Model	Development MAE (min)	Test MAE (min)	MedAE (min)	RMSE (min)	Coverage (%)	Mean width (min)	Interval score (min)
M0 (preferred)	95.77	49.25	37.00	78.99	91.5	308.00	333.59
M0R	95.75	48.30	34.15	79.97	87.8	248.30	288.80
M1	94.91	47.66	33.09	78.89	91.6	312.65	335.92
M2	95.05	48.02	33.47	79.10	91.4	307.67	332.06

M0 = historical benchmark; M0R = recent 90-day benchmark; M1 = quantile regression; M2 = LightGBM. MAE = mean absolute error; MedAE = median absolute error; RMSE = root mean squared error

Recovery medians were below pre-COVID within all four age bands and both recorded sex categories (Supplementary Figure S1). Acuity patterns were less uniform: Resuscitation medians were 18.0, 38.0 and 20.0 minutes, leaving recovery 2.0 minutes above pre-COVID in that category. Emergency medians were 68.0, 53.0 and 23.0 (Supplementary Figure S2). Within recovery, Non Urgent increased from 46.0 minutes before the recording boundary to 59.8 thereafter, exceeding its pre-COVID median of 56.0 (Supplementary Tables S5–S6).

Sensitivity and supporting facility results

Primary phase medians were unchanged to one decimal place when zero or positive Error-coded intervals were included, long intervals or chronology flags were excluded, or monthly quality screens were applied. Tail-sensitive measures varied more. Of 85 stays exceeding 72 hours, 82 preceded COVID. Removing these reduced the pre-COVID mean from 143.66 to 130.70 minutes while leaving its median at 63.0. Omitting March–May 2020 increased the COVID 90th percentile from 294.0 to 302.0 minutes despite an unchanged median. Alternative calendar-year groups yielded medians of 64.0, 62.0 and 51.0 minutes. The 95% monthly screen and omission of March–May 2020 selected the same records and did not provide independent checks.

The eight selected facility codes had unequal observation windows and source labels spanning Balqa, Zarqa and Madaba (Supplementary Figure S3). Supporting results were reported separately. At facility 124, a common 2018–June 2024 window changed the pre-COVID median from 77.0 to 48.0 minutes, reversing the COVID-minus-pre-COVID contrast from –22.0 to +7.0. At facility 4, including positive Error records changed the COVID

median from 52.0 to 62.0, reversing the recovery-minus-COVID contrast from +3.02 to −6.98 minutes.

Development selection and later prediction

Development, calibration and testing contained 168,836, 12,274 and 191,497 encounters. Regression (M1) had the smallest pooled development mean absolute error (MAE), 94.91 minutes. The historical benchmark (M0) was preferred because its 95.77-minute MAE met the fixed 1% simplicity tolerance. The final fitted M0 predicted 64 minutes for every test encounter. Selection chose an unpenalised M1 fit and M2 with seven leaves, minimum child size 100 and L2 penalty 1. Test MAE was 49.25 minutes for M0, 48.30 for the recent benchmark (M0R), 47.66 for M1 and 48.02 for LightGBM (M2; Table 5). Relative to M0, test MAE was 1.59 minutes lower for M1 and 1.24 minutes lower for M2. The calendar-only tree diagnostic had MAE 48.13 minutes; adding age and sex reduced this by 0.115 minutes.

Prediction intervals and heterogeneity

All seven evaluated configurations required zero calibration expansion, leaving the repaired intervals unchanged; calibration-sample coverage ranged from 82.0% to 86.9%. M0's nominal 80% interval covered 91.5% of test observations with a constant width of 308.0 minutes. M1 and M2 covered 91.6% and 91.4%, with mean widths of 312.6 and 307.7 minutes. M0R had the narrowest main-model interval, 248.3 minutes, and the lowest interval score, 288.80, despite a higher MAE than M. Observed outcomes fell below the fitted median in 58.6% of M1 and 58.7% of M2 predictions, and below their fitted 90th quantiles in approximately 98.2% (Supplementary Figure S4).

Temporal and subgroup checks showed differences hidden by aggregate performance (Figure 4). M0 MAE was 51.93 minutes before the recording boundary and 47.62 thereafter; coverage was 89.1% and 92.9%. Quarterly M0 MAE ranged from 43.58 minutes in 2025 Q2

to 63.37 in 2023 Q3. Among 1,035 Resuscitation encounters, coverage was only 68.7% for M0, 71.8% for M1 and 71.2% for M2. Most misses were below the lower limit. Emergency coverage was 79.0% for M0 but 90.8% for M1 and 89.0% for M2. Predictive gains were not uniform across age: among 10,721 encounters aged ≥ 65 years, M1 and M2 MAEs were 50.51 and 50.72 minutes, compared with 50.24 for M0.

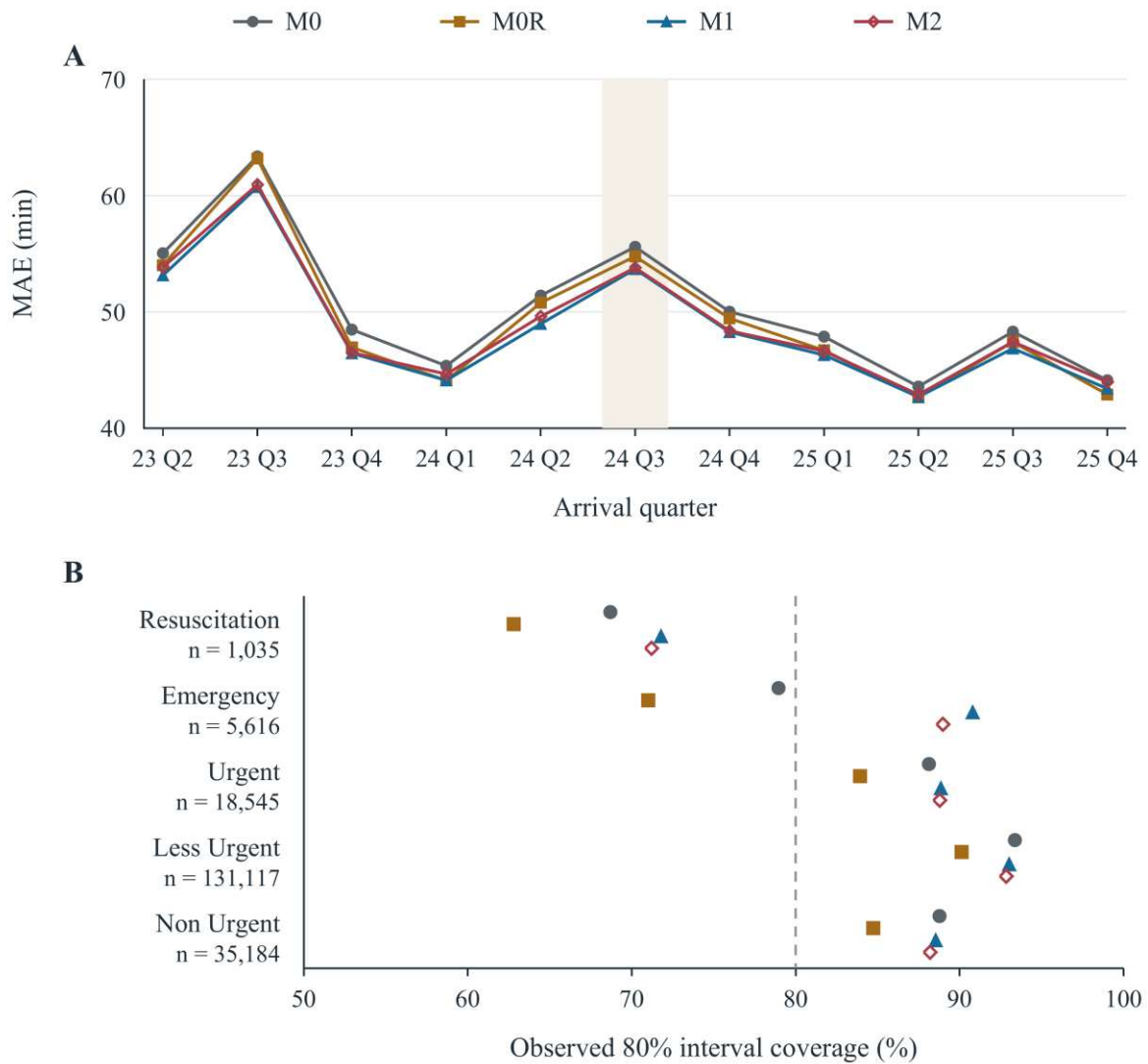


Figure 4: Temporal prediction error and interval coverage within recorded acuity groups

A: test MAE by quarter; 23–25 denote 2023–2025. The first quarter begins 6 May 2023; shading identifies 2024 Q3, which spans the recording boundary. B: inclusive coverage with group counts; the dashed line marks 80%. M0 = historical benchmark; M0R = recent benchmark; M1 = regression; M2 = LightGBM.

Prediction sensitivity checks

The main-model MAE ordering persisted in both patient-disjoint subsets, after omitting the final test quarter, and after adding nine positive Error-coded encounters. The two patient exclusions retained 63,049 and 60,553 encounters. Removing development stays above 72 hours changed M1 MAE by -0.011 minutes and M2 by $+0.005$ minutes on the unchanged full test cohort. These refits remained sensitivities and did not replace the selected model. No test encounters had missing age or unknown sex, so prediction under those missing-input conditions was not evaluated (Supplementary Tables S7–S8).

Discussion

Principal findings

PHASE-ED found shorter recorded stays during recovery, with a larger change in the upper tail than at the median (Figure 5). In the primary facility, the median declined from 63.0 minutes before COVID to 51.0 during recovery, while the 90th percentile declined from 317.0 to 156.6 minutes. The median differences persisted across several eligibility checks, and the lower recovery median was present before the July 2024 precision transition. It is evidence of a temporal change in the recorded distribution. Whether that change represents faster care, different patients or altered documentation remains unresolved. More complex models produced only small reductions in prediction error, and reassuring overall interval coverage concealed undercoverage in some recorded acuity groups.

Interpretation alongside emergency care research

Jordanian research already documents public-hospital emergency department use and duration categories (Rawshdeh et al., 2025). PHASE-ED extends this evidence through explicit recording-period checks, longer follow-up and later-period prediction assessment; it does not establish the first local analysis of length of stay. Faris et al. (2026) reported shorter mean stays after lockdown alongside lower acuity. The direction resembles PHASE-ED, where the Less Urgent category increased from 46.05% to 68.47%. However, different phase

boundaries and PHASE-ED's unverified initial-acuity timing prevent interpreting the similarity as an equivalent change in clinical severity.

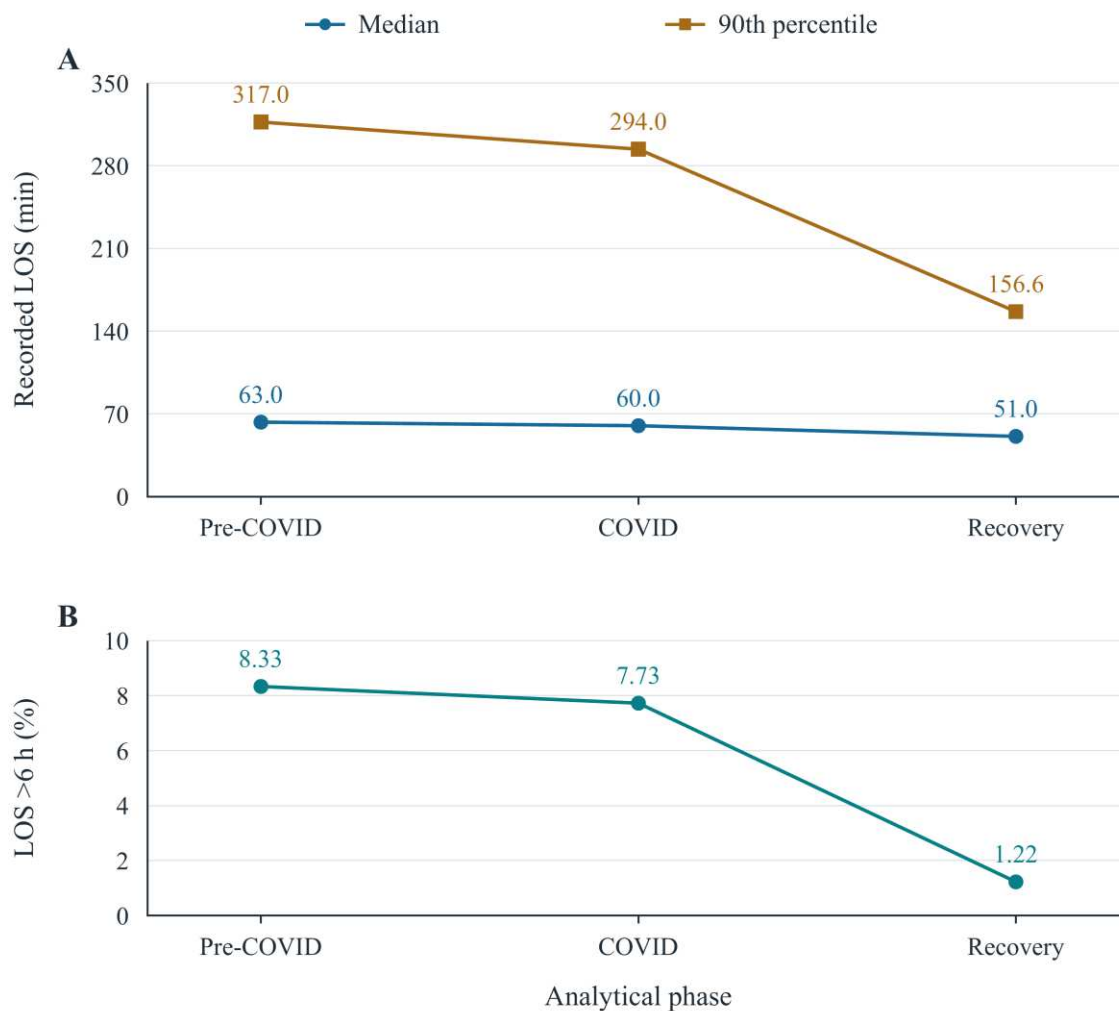


Figure 5: Changes in the centre and upper tail of recorded emergency department stay

A: encounter-level median and 90th percentile. B: proportion exceeding 6 hours. The primary facility contributed 72,840 pre-COVID, 117,681 COVID and 191,497 recovery encounters.

International studies also show heterogeneous service trajectories. Across 111 US emergency departments, admitted-patient median stay increased from 5.2 hours in 2019 to 6.9 in 2022 despite persistently lower visit volumes (Oskvarek et al., 2023). Among older US adults, the proportion staying longer than eight hours rose from 12% in 2017 to 20% in 2024 (Haimovich et al., 2025). These findings are contextual comparisons (Table 6). Their populations, thresholds and endpoints differ from PHASE-ED's all-encounter distributions.

Table 6: Selected comparisons of emergency department duration across settings

Study and setting	Observation period and population	Relevant finding	Comparison with PHASE-ED
Rawshdeh et al. (2025) Jordanian public hospitals	January 2018–November 2023; routine emergency department records	Reports utilisation, admission and duration categories.	Local setting precedent; does not verify the present facility roster or timestamp definitions.
Faris et al. (2026) Beirut, Lebanon	February 2020–April 2022; 69,870 adult and paediatric records	Mean stay decreased from 330.5 to 319.2 minutes after lockdown, alongside lower acuity.	Different pandemic boundaries; mean rather than median. Similar direction does not establish a shared mechanism.
Haimovich et al. (2025) United States	2017–2024; adults aged ≥ 65 years	Proportion with total stay >8 hours increased from 12% to 20%.	Older patients and an 8-hour threshold; not directly comparable to an all-age 6-hour outcome.
Oskvarek et al. (2023) 111 US emergency departments	March 2020–August 2022 versus corresponding 2019 months	Admitted-patient median stay was 5.2 hours in 2019 and 6.9 in 2022 despite lower visit volumes.	Admitted-patient, site-aggregated summaries; follow-up predates PHASE-ED recovery.

Direct observation has shown that electronic timestamp accuracy can change following implementation of a health-record system (Ward et al., 2013). In PHASE-ED, recovery medians of 50.0 minutes before the precision boundary and 51.2 thereafter indicate that the lower median was not confined to the later recording period. They do not rule out earlier recording changes, incomplete capture or composition effects. Nor do stable medians under cleaning sensitivities validate the clinical meaning of arrival and departure. The marked change in unknown disposition further cautions against assuming an unchanged recording process.

What the prediction analysis adds

Relative to the historical benchmark, quantile regression and boosted trees reduced test mean absolute error (MAE) by 1.59 and 1.24 minutes. The development rule retained the benchmark because its error was within 1% of the smallest value. Neither model improved MAE in the group aged 65 years or older. Adding age and sex to the calendar-only tree reduced overall MAE by only 0.115 minutes; this small predictive increment does not

identify causal determinants. Abaddi (2025b) compared neural-network forecasts with conventional methods for dairy sales, providing a cross-domain example of benchmark-based evaluation. The methodological connection is the need to assess added complexity against simpler alternatives; differences in outcomes and time scales preclude direct performance comparisons with PHASE-ED.

Other emergency department studies address different prediction tasks. Canellas et al. (2024) classified prolonged process intervals at successive operational landmarks; Song et al. (2025) predicted duration categories after triage. Wu et al. (2025) included completed waiting time, which by definition is unavailable at arrival. Their outcomes and information sets preclude numerical accuracy rankings against PHASE-ED. Our findings concern retrospective encounter-level duration prediction, not forecasting of aggregate demand, and do not show that simpler models will remain preferable after richer prospectively available information is added.

The historical benchmark's nominal 80% interval covered 91.5% overall but spanned 308.0 minutes; coverage was only 68.7% within recorded Resuscitation. Most subgroup misses fell below the lower limit, so the problem was not simply failure to capture exceptionally long stays. Overall coverage does not assure subgroup coverage (Gibbs et al., 2025). All calibration expansions were zero: the expansion-only procedure left the existing intervals unchanged and could not narrow them. Temporal change and repeated attendance also preclude a conventional exchangeability-based coverage guarantee (Barber et al., 2023). The recent benchmark offered narrower intervals and a lower interval score despite greater point error than regression, showing why MAE alone is insufficient (Supplementary Figure S1). Evaluation should therefore begin with a defined operational decision and assess contemporary calibration, subgroup performance and utility before any workflow intervention (Riley et al., 2024).

Implications for service measurement and practice

Asplin et al.'s (2003) input, throughput and output framework helps locate the missing information. PHASE-ED measures a combined interval; it cannot distinguish consultation delay, investigation time or waiting for an inpatient bed. A Jordanian improvement project reported a reduction in mean consultation time from 91 to 65 minutes, illustrating the value of examining a specific process (Al-Na'seh et al., 2024). That uncontrolled result is not a treatment effect transferable to this cohort. The next operational step is to validate event definitions and add clinician-contact, disposition-decision and physical-departure times, alongside staffing and bed availability. Managers could then monitor the median and upper tail with recording-quality indicators, while auditing specific delays before changing staffing or discharge practices.

Abaddi (2025a) combined time-and-motion measurement with simulation to evaluate manufacturing workflows. For PHASE-ED, this suggests a methodological extension: validate clinical timestamps, then simulate specific emergency department process changes before prospective evaluation.

Strengths, limitations and next validation

The eight-year primary series, transparent exclusions, preserved repeat visits and outcome-tail checks strengthen the description of available records. Patient-disjoint sensitivity tests also reduced one source of overlap in predictive evaluation. Nevertheless, the primary inference concerns one facility; seven supporting codes have unequal windows and do not form a representative national panel. At facility 124, restricting the observation window reversed the COVID-versus-pre-COVID contrast.

Current demographic values do not establish their historical availability. Comparisons were unadjusted; repeated attendance, secular trends, selection and changing composition remain relevant. No confidence intervals or causal effects were estimated. Low missingness

in the selected cohort does not establish completeness of the original extraction. Model development followed inspection of historical test-outcome distributions, and was not preregistered or evaluated prospectively. Performance with missing age or unknown sex was not tested because those strata were empty. Future validation should use a new period and independently verified facilities, with subgroup calibration and decision-specific utility assessed before implementation.

Conclusion

Recorded emergency department stays shortened during recovery at the primary facility, particularly in the upper tail. The evidence supports further investigation of this change. Sparse-variable prediction offered modest gains, while interval performance remained uneven. Better event validation and prospective testing are necessary before these findings guide operational decisions.

Abbreviations

AI — Artificial intelligence

AUBMC — American University of Beirut Medical Center

COVID-19 (COVID) — Coronavirus disease 2019

CPU — Central processing unit

ED — Emergency department

GiB — Gibibyte

GPU — Graphics processing unit

IPM — Interior-point method

IS — Interval score

JSON — JavaScript Object Notation

KAUH — King Abdullah University Hospital

LightGBM — Light Gradient Boosting Machine

LOS — Length of stay

M0 — Historical empirical quantile benchmark

M0R — Recent 90-day empirical quantile benchmark

M1 — Spline quantile regression

M2 — LightGBM quantile tree model

MAE — Mean absolute error

MedAE — Median absolute error

ORCID — Open Researcher and Contributor ID

P90 — 90th percentile

PHASE-ED — Study project name

Q1–Q3 — First to third quartiles

RECORD — REporting of studies Conducted using Observational Routinely-collected health Data

RMSE — Root mean squared error

SD — Standard deviation

SHA-256 — Secure Hash Algorithm with a 256-bit digest

TRIPOD+AI — Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis plus Artificial Intelligence

US — United States

UTC — Coordinated Universal Time

WHO — World Health Organization

Declarations

Ethics approval and consent to participate

N/A

Availability of data and materials

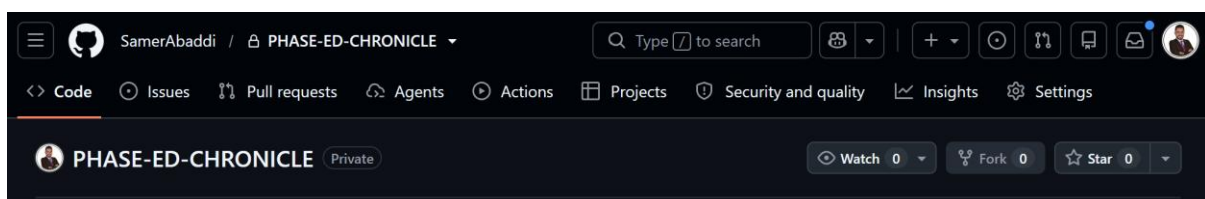
Patient-level records, fitted model objects and individual predictions are not publicly available and remain in the private Google Cloud research environment. Aggregate analysis tables, comparisons of model performance, and figure-data materials are included in the study and the supplementary materials. Any external disclosure or access requires authorisation from the relevant data custodian.

Three videos were shared with this submission. Video 1 shows the private repository that contains 56 related files: cleaning SQL, modelling and figure code, tests, software versions, and reproducibility documentation. Video 2 shows a sample of the log schemas, cleaning summary, and model comparison results. Video 3 is a summarized research notebook/ website that reviews the work.

Code availability

The source code is uploaded to GitHub under a private repository titled: PHASE-ED-CHRONICLE.

Exceptional access to the GitHub repository can be granted at the Editor-in-Chief's request. Post-acceptance, the repository is intended to be public and linked to the article DOI's. Code files can be sent to the journal as well.



Funding

This research received no funding.

Competing interests

The author declares no competing interests.

Author contributions

Samer Abaddi: conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing – original draft, writing – review and editing, visualization, supervision, and project administration. Funding acquisition was not applicable. Samer Abaddi is the sole and corresponding author and is responsible for the final manuscript.

Use of generative artificial intelligence

AI assisted with proofreading. The author takes full responsibility for the final publication.

Acknowledgements

N/A

References

- Abaddi, S. (2025a). Optimizing double glazing through time and work measurements and simulation. *Journal of Simulation*, 19(5), 555–577.
<https://doi.org/10.1080/17477778.2024.2446749>
- Abaddi, S. (2025b). Optimizing sales forecasting in dairy industry using neural networks. *Jordan Journal of Mechanical and Industrial Engineering*, 19(2), 415–426.
<https://doi.org/10.59038/jjmie/190213>
- Abaddi, S. (2026). The optimization and load forecasting of power distribution units in data halls. *Applied Operations and Analytics*, 2(1), 1–25.
<https://doi.org/10.1080/29966892.2026.2649444>
- Al-Na'seh, M. H., Elheet, A. M., Alhayek, A. M., Sabri, A. T., & Al Owaidat, A. K. (2024). Optimizing emergency department length of stay and quality of care: A quality improvement project. *Cureus*, 16(10), Article e71989.
<https://doi.org/10.7759/cureus.71989>
- Al-Qudah, M., Al-Shaikh, A., Haddad, H., Alolayyan, M., Zawaneh, Y., Matalaka, M., & Matalaka, I. (2022). Impact of COVID-19 exclusive allocation strategy on quality of healthcare: A study from Jordan, 2020. *Health Security*, 20(3), 193–202.
<https://doi.org/10.1089/hs.2021.0146>
- Asplin, B. R., Magid, D. J., Rhodes, K. V., Solberg, L. I., Lurie, N., & Camargo, C. A., Jr. (2003). A conceptual model of emergency department crowding. *Annals of Emergency Medicine*, 42(2), 173–180. <https://doi.org/10.1067/mem.2003.302>
- Barber, R. F., Candès, E. J., Ramdas, A., & Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2), 816–845.
<https://doi.org/10.1214/23-AOS2276>

- Benchimol, E. I., Smeeth, L., Guttman, A., Harron, K., Moher, D., Petersen, I., Sørensen, H. T., von Elm, E., Langan, S. M., & RECORD Working Committee. (2015). The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLOS Medicine*, 12(10), Article e1001885. <https://doi.org/10.1371/journal.pmed.1001885>
- Canellas, M. M., Kotkowski, K. A., Pachamano, D. A., Perakis, G., Rezek, M. A., Skali Lami, O., & Tsiourvas, A. (2024). A granular view of emergency department length of stay: Improving predictive power and extracting real-time, actionable insights. *Annals of Emergency Medicine*, 84(4), 386–398. <https://doi.org/10.1016/j.annemergmed.2024.02.004>
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, Article e078378. <https://doi.org/10.1136/bmj-2023-078378>
- Department of Statistics. (2026). *Statistical yearbook of Jordan 2024*. https://dosweb.dos.gov.jo/databank/yearbook/YearBook_2024.pdf
- Electronic Health Solutions. (2026). *Hakeem program*. <https://ehs.com.jo/hakeem-program>
- Farimani, R. M., Karim, H., Atashi, A., Tohidinezhad, F., Bahaadini, K., Abu-Hanna, A., & Eslami, S. (2024). Models to predict length of stay in the emergency department: A systematic literature review and appraisal. *BMC Emergency Medicine*, 24, Article 54. <https://doi.org/10.1186/s12873-024-00965-4>

- Faris, N., Hamdar, L., Tamim, H., Makki, M., & Hitti, E. (2026). Emergency department visits and disease patterns during and after the COVID-19 lockdown. *PLOS ONE*, 21(6), Article e0352275. <https://doi.org/10.1371/journal.pone.0352275>
- Gibbs, I., Cherian, J. J., & Candès, E. J. (2025). Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4), 1100–1126. <https://doi.org/10.1093/jrsss/qkaf008>
- Haimovich, A. D., Berry, S. D., & Landon, B. E. (2025). Prolonged emergency department stays for older US adults. *JAMA Internal Medicine*, 185(8), 1037–1038. <https://doi.org/10.1001/jamainternmed.2025.2006>
- Jones, S., Moulton, C., Swift, S., Molyneux, P., Black, S., Mason, N., Oakley, R., & Mann, C. (2022). Association between delays to patient admission from the emergency department and all-cause 30-day mortality. *Emergency Medicine Journal*, 39(3), 168–173. <https://doi.org/10.1136/emered-2021-211572>
- Oskvarek, J. J., Zocchi, M. S., Black, B. S., Celedon, P., Leubitz, A., Moghtaderi, A., Nikolla, D. A., Rahman, N., & Pines, J. M. (2023). Emergency department volume, severity, and crowding since the onset of the coronavirus disease 2019 pandemic. *Annals of Emergency Medicine*, 82(6), 650–660. <https://doi.org/10.1016/j.annemergmed.2023.07.024>
- Rawshdeh, M., Obeidat, S., Alwadi, M., & Alrawashdeh, A. (2025). Characteristics of emergency department visits in Jordanian public hospitals. *International Journal of Emergency Medicine*, 18, Article 166. <https://doi.org/10.1186/s12245-025-00952-x>
- Rawshdeh, M., Obeidat, S., Alwadi, M., & Alrawashdeh, A. (2025). Characteristics of emergency department visits in Jordanian public hospitals. *International Journal of Emergency Medicine*, 18, Article 166. <https://doi.org/10.1186/s12245-025-00952-x>

- Riley, R. D., Archer, L., Snell, K. I. E., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., & Collins, G. S. (2024). Evaluation of clinical prediction models (part 2): How to undertake an external validation study. *BMJ*, 384, Article e074820. <https://doi.org/10.1136/bmj-2023-074820>
- Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 3543–3553). Curran Associates. <https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html>
- Song, L., Aickelin, U., Fazio, T. N., Sharma, A., Kouhounestani, M., Plumb, S., & Putland, M. J. (2025). Developing interpretable machine learning models to predict length of stay and disposition decision for adult patients in emergency departments. *BMJ Health & Care Informatics*, 32(1), Article e101152. <https://doi.org/10.1136/bmjhci-2024-101152>
- Ward, M. J., Froehle, C. M., Hart, K. W., & Lindsell, C. J. (2013). Operational data integrity during electronic health record implementation in the ED. *The American Journal of Emergency Medicine*, 31(7), 1029–1033. <https://doi.org/10.1016/j.ajem.2013.03.027>
- World Health Organization. (2021). *WHO supports generating evidence for decision-making in Jordan during COVID-19*. <https://www.who.int/about/accountability/results/who-results-report-2020-mtr/country-story/2020/jordan>

- World Health Organization. (2023, May 5). *Statement on the fifteenth meeting of the IHR (2005) Emergency Committee on the COVID-19 pandemic*.
[https://www.who.int/news/item/05-05-2023-statement-on-the-fifteenth-meeting-of-the-international-health-regulations-\(2005\)-emergency-committee-regarding-the-coronavirus-disease-\(covid-19\)-pandemic](https://www.who.int/news/item/05-05-2023-statement-on-the-fifteenth-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-coronavirus-disease-(covid-19)-pandemic)
- Wu, W., Li, M., Jiang, H., Sun, M., Zhu, Y., Zhu, G., Li, Y., Li, Y., Mo, J., Chen, X., & Mao, H. (2025). Development of an emergency department length-of-stay prediction model based on machine learning. *World Journal of Emergency Medicine*, 16(3), 220–224.
<https://doi.org/10.5847/wjem.j.1920-8642.2025.048>

Supplementary materials

Table S1: Facility observation windows and distinct analytical roles

Code	Role	Source governorate	Arrival window	Arrivals	Eligible LOS
273	Primary	Balqa	2018-01-01 to 2025-12-31	382,781	382,018
124	Supporting	Zarqa	2017-01-01 to 2024-06-30	2,137,347	2,107,009
4	Exploratory	Balqa	2018-05-01 to 2025-12-31 First observed: 2018-05-02	1,111,690	1,013,026
192	Supporting	Madaba	2023-03-01 to 2025-02-28	224,859	224,531
305	Supporting	Balqa	2020-10-01 to 2025-12-31	1,277,468	1,243,420
399	Recovery only	Balqa	2024-09-01 to 2025-12-31	108,842	108,287
152	Available-period description	Zarqa	2017-01-01 to 2025-12-31	1,966,314	1,829,219
272	Available-period description	Madaba	2017-01-01 to 2025-12-31 First observed: 2018-01-07	135,209	134,803

Note. Dates are inclusive. Rule dates differ from the first observed record for codes 4 and 272. Codes 210 and 268 supported arrival-quality assessment only; 156 and 288 were trace-code audits.

Table S2: Definitions of sensitivity analyses and recording checks

Analysis	Check	Executed rule
Descriptive	Outcome exclusions	Separately include positive Error-coded records, or include zero LOS while excluding Error. Negative or missing durations are never repaired or reclassified as observed positive outcomes.
Descriptive	Upper-tail restrictions	Restrict otherwise-eligible LOS separately to ≤ 72 hours, ≤ 7 days and ≤ 30 days. These are nested sensitivity populations; the primary analysis retains all positive durations.
Descriptive	Chronology and age	Exclude BirthAfterArrival or DeathBeforeArrival flags. DeathBeforeBirth was retained as an audit flag but was not an additional exclusion predicate. A separate plausible birth-derived-age diagnostic examined numeric age and age-band differences.
Descriptive	Alternative periods	Compare complete calendar groups 2018–2019, 2020–2022 and 2024–2025, omitting 2023. Separately omit March–May 2020 under the original phase boundaries.
Descriptive	Monthly quality screens	At each of the 90% and 95% cutoffs, require both active arrival days/calendar days and positive-duration records/all arrivals to meet the cutoff. Positive duration includes Error for the quality fraction; final LOS eligibility still excludes Error. The 95% screen and March–May 2020 omission selected identical records and are not independent checks.
Descriptive	Supporting windows	Analyse each facility and role separately with main and Error-inclusive summaries. Facility 124 also excludes 2018 in a sensitivity; facilities 124

		and 273 use a common 1 January 2018–30 June 2024 window in separate analyses. Facility 4 has an additional Error-inclusive monthly series.
Descriptive	Precision and subgroups	Compare exact, floored and rounded-minute phase medians on the same eligible records; inspect negative-duration bands and departure precision. Describe age, sex and recorded acuity by phase and before/from 23 July 2024. Unknown age/acuity subgroup LOS estimates are suppressed to counts only, with records retained in totals.
Prediction	Unseen patients	Exclude patients in any development/calibration arrival, including outcome-ineligible arrivals: 63,049 test encounters from 27,564 patients. The stricter check excludes anyone with any facility-273 arrival during 1 January 2018–5 May 2023, including both gaps: 60,553 encounters from 26,846 patients. Neither establishes no attendance before 2018.
Prediction	Terminal period and Error	Omit October–December 2025 without refitting: 167,153 test encounters remain. Separately add 9 otherwise-positive Error-coded encounters to the frozen primary test, giving 191,506; this changes the evaluated population.
Prediction	Training upper tail	Refit M1 and M2 using 168,751 development encounters (54,772 patients) with $LOS \leq 4,320$ minutes, retaining the main-selected hyperparameters. Fit preprocessing on restricted development only. Recalibrate on the unchanged calibration cohort and evaluate the unchanged full primary test; do not retune or replace the selected primary family.
Prediction	Calibration and diagnostic bins	Evaluate clipped fitted quantiles before envelope repair for pinball loss. Use inclusive interval coverage, separate lower/upper miss rates, widths and interval score. Prediction-bin cutpoints are pooled development out-of-fold deciles, with duplicate edges collapsed and closed-right bins; calibration and test outcomes do not define bins. Sensitivity refits reuse their main-model bin edges.

Table S3: Model specifications and complete estimator settings

Model or operation	Candidates or fixed specification	Selected or executed setting
M0: historical benchmark	Empirical quantiles 0.10, 0.50, 0.90 of training outcomes; linear interpolation.	Final development: 168,836 encounters.
M0R: recent benchmark	Same empirical quantiles using arrivals in the final 90 calendar days of each training window.	Final window: 2022-09-02 to 2022-11-30; 9,318 encounters.
M1: QuantileRegressor	L1 alpha candidates: 0, 0.01, 0.1, 1; median quantile 0.50 for tuning.	Selected alpha=0 for main and 72-hour training-tail refit.
M1: all remaining settings	fit_intercept=True; solver=highs-ipm; solver_options={threads: 4}.	Separate final quantile=0.10, 0.50 and 0.90 fits.
M2: tuned grid	num_leaves: 7, 15, 31;	Main: 7 leaves, minimum child size

	min_child_samples: 100, 500; reg_lambda: 0, 1.	100, L2=1. Calendar-only: 7, 100, L2=0.
M2: boosting and loss	boosting_type=gbd; objective=quantile; learning_rate=0.05; n_estimators=500.	alpha=0.50 for tuning; separate final alpha=0.10, 0.50, 0.90.
M2: split and depth defaults	max_depth=-1; min_child_weight=0.001; min_split_gain=0; max_bin=255.	Unchanged for main, calendar-only and training-tail fits.
M2: sampling and penalties	reg_alpha=0; subsample=1; subsample_freq=0; colsample_bytree=1; subsample_for_bin=200000.	reg_lambda is the selected value above.
M2: execution and remaining defaults	n_jobs=4; random_state=20261012; deterministic=True; force_col_wise=True.	class_weight=None; importance_type=split; verbosity=-1.
Within-family selection	Minimum pooled out-of-fold MAE of nonnegative-clipped median predictions.	Three validation windows; 98,175 total observations.
Exact parameter ties	M1: prefer higher alpha. M2: fewer leaves, then higher min_child_samples, then higher reg_lambda.	Applied only to exactly equal pooled MAE values.
Across-family preference	Among M0/M0R/M1/M2 with pooled MAE ≤ 1.01 times the smallest main- family MAE, prefer M0, then M0R, then M1, then M2.	M0 preferred; M1 had the smallest development MAE.
Final quantile fits	Reuse median-selected settings at quantiles 0.10 and 0.90, without tail- specific tuning.	Main full-development models plus calendar diagnostic; restricted-training M1/M2 use the same selected settings.
Fit accounting	Inner median fits: 4x3 for M1; 12x3 for M2; 12x3 for calendar-only M2.	84 inner estimator fits plus 15 final estimator fits = 99.

Table S4: Model specifications and complete estimator settings

Item	Recorded value
Training environment	Google Cloud Shell; Linux 6.6.157+; x86_64; glibc 2.39.
Python	3.12.3; GCC 13.3.0.
CPU and thread budget	4 visible CPUs; 4 threads per fit; sequential candidate fits.
Memory and hardware limits	Host MemTotal: 16,382,832 kB (approximately 15.6 GiB); MemAvailable at capture: 15,569,252 kB.
Randomness	Seed/random_state: 20261012; LightGBM deterministic=True and force_col_wise=True.
Numerical packages	NumPy 2.4.6; SciPy 1.17.1; pandas 3.0.5.
Modelling packages	scikit-learn 1.9.0; LightGBM 4.7.0; joblib 1.5.3; threadpoolctl 3.6.0.
Data-access packages	google-cloud-bigquery 3.43.0; PyArrow 26.0.0; db-dtypes 1.7.2;

	google-auth 2.56.3.
Source/data verification	22 source-contract checks; 164 record-level checks; source metadata and content digests unchanged before/after preparation.
Synthetic implementation tests	39 tests passed, with no failures, errors or skips.
Numerical amendment	2026-10-12 16:49:46.919183 UTC: M1 solver changed from highs to highs-ipm before study fitting.
Amendment checks	12 parameter/quantile cells on 2,000 synthetic rows; maximum objective difference 3.55e-15. Full-size IPM synthetic fit: 168,836 rows, 391.055 seconds.
Manifest freeze	2026-10-12 16:50:05.880833 UTC.
Executed code: phase_ed_models.py	f570e9e3a18f4579e7a5636e12d7a2223b2d229d246d5cf057deb2f01b94ff83
Executed code: phase_ed_evaluation.py	c23d63691a383674c55aacc47c2e339b26c22c8955ff1d8ce4ccf48071abb922
Executed code: phase_ed_data.py	fe9e4f6a2988716d3453f88a9a7554cc278072f484f7aeadb28e7df362d07d2
Executed code: run_private_training.py	f59f90361589c4187a1394d1ff074901d38b9cd24a31044f955ba7928b6a06ec
Executed code: private_training_extract.sql	e536d20b5f7424555514d61079fcb86d2b77c47f7557140573a5f8f76884d487
Selection receipt	2026-10-12T17:06:51.361969+00:00 SHA-256: 936a65c5f979cd8706cef44d696f40b01b33039e2dc569a02e6aba5d61f9b3d8
Final fits receipt	2026-10-12T17:27:15.145084+00:00 SHA-256: 9a2549fdb4d4a9502615d48860b73cdccbe03c1d14e05f3418bd629a3ca31c47
Calibration receipt	2026-10-12T17:27:18.381930+00:00 SHA-256: 609e67b0d0f2a858d3f039370b470bc449f44c60cf6bc8a699d7fc0f1846b727
Evaluation receipt	2026-10-12T17:27:55.987150+00:00 SHA-256: 83a287df047322ac2579575cbc021ec24c0b63add352560ddfb3a887c678ded5
Original scientific protocol SHA-256	66237c946a12fdf628a58c88afedb704d18e54f275d633d54f148843c932da27
Effective scientific protocol SHA-256	3af1e68941229c8c74bf79167ef49bfba1be54d40d83f133c66822c5e0f09464

Table S5: Primary recorded LOS by age, sex and recorded acuity

Stratum	Pre-COVID	COVID	Recovery
Age: <18	n = 23,650 61.0 (21.3–139.8) P90 307.0; >6 h 1,882 (7.96%)	n = 36,109 56.0 (24.0–126.0) P90 283.0; >6 h 2,731 (7.56%)	n = 65,597 49.4 (24.0–90.0) P90 151.8; >6 h 767 (1.17%)
Age: 18-39	n = 28,893 59.0 (19.0–142.0) P90 308.0; >6 h 2,337 (8.09%)	n = 46,421 58.0 (24.0–132.0) P90 300.0; >6 h 3,670 (7.91%)	n = 71,462 49.4 (23.0–91.7) P90 155.0; >6 h 939 (1.31%)
Age: 40-64	n = 15,449 71.0 (24.0–160.0) P90 334.0; >6 h 1,352 (8.75%)	n = 27,560 64.0 (27.0–138.0) P90 298.0; >6 h 2,144 (7.78%)	n = 43,717 53.0 (25.0–97.5) P90 161.1; >6 h 534 (1.22%)
Age: 65+	n = 4,844 80.0 (26.0–172.0) P90 370.0; >6 h 499 (10.30%)	n = 7,591 70.0 (29.0–139.0) P90 287.0; >6 h 547 (7.21%)	n = 10,721 60.0 (27.8–104.9) P90 166.0; >6 h 103 (0.96%)
Age: Unknown	n = 4; counts only	n = 0; not estimated	n = 0; not estimated
Sex: Female	n = 30,227 67.0 (23.0–152.0) P90 331.0; >6 h 2,652 (8.77%)	n = 49,973 61.0 (26.0–131.0) P90 285.0; >6 h 3,787 (7.58%)	n = 79,227 51.0 (24.0–92.7) P90 153.5; >6 h 953 (1.20%)
Sex: Male	n = 42,613 61.0 (20.0–144.0) P90 308.0; >6 h 3,418 (8.02%)	n = 67,708 59.0 (24.0–133.0) P90 301.0; >6 h 5,305 (7.84%)	n = 112,270 50.8 (24.0–93.8) P90 158.2; >6 h 1,390 (1.24%)
Acuity: Resuscitation	n = 499 18.0 (5.0–92.5) P90 296.8; >6 h 42 (8.42%)	n = 1,792 38.0 (11.0–153.0) P90 382.9; >6 h 211 (11.77%)	n = 1,035 20.0 (6.9–64.5) P90 141.4; >6 h 23 (2.22%)
Acuity: Emergency	n = 2,808 68.0 (20.0–168.0) P90 348.0; >6 h 261 (9.29%)	n = 11,261 53.0 (16.0–180.0) P90 396.0; >6 h 1,252 (11.12%)	n = 5,616 23.0 (10.0–63.9) P90 177.2; >6 h 160 (2.85%)
Acuity: Urgent	n = 11,012 86.0 (29.0–175.0) P90 315.0; >6 h 886 (8.05%)	n = 13,744 85.0 (34.0–210.0) P90 434.0; >6 h 1,877 (13.66%)	n = 18,545 52.0 (22.0–99.1) P90 171.0; >6 h 455 (2.45%)
Acuity: Less Urgent	n = 33,546 63.0 (23.0–148.0) P90 342.0; >6 h 3,116 (9.29%)	n = 63,364 56.0 (25.0–112.0) P90 205.0; >6 h 2,731 (4.31%)	n = 131,117 51.0 (25.9–90.0) P90 146.0; >6 h 1,102 (0.84%)
Acuity: Non Urgent	n = 24,971 56.0 (17.0–131.0) P90 284.0; >6 h 1,765 (7.07%)	n = 27,520 62.0 (26.0–151.0) P90 385.0; >6 h 3,021 (10.98%)	n = 35,184 55.0 (23.2–109.0) P90 182.3; >6 h 603 (1.71%)

Acuity: Unknown	n = 4; counts only	n = 0; not estimated	n = 0; not estimated
--------------------	--------------------	----------------------	----------------------

Table S6: Recorded subgroup patterns before and from the observed recording boundary

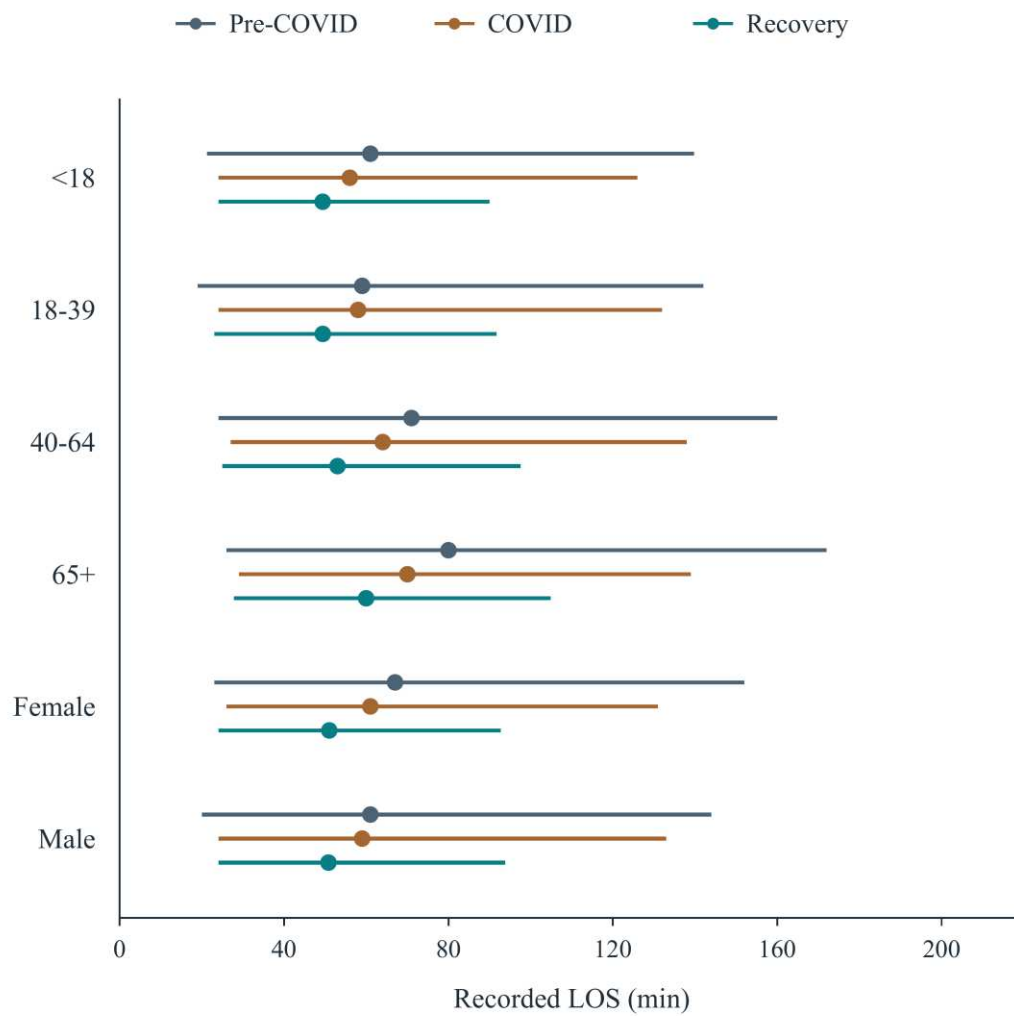
Stratum	Recovery before 23 July 2024	Recovery from 23 July 2024
Age: <18	n = 24,398 48.0 (21.0–91.0) P90 153.0; >6 h 345 (1.41%)	n = 41,199 50.0 (25.3–89.6) P90 151.2; >6 h 422 (1.02%)
Age: 18-39	n = 27,519 48.0 (21.0–92.0) P90 158.0; >6 h 428 (1.56%)	n = 43,943 50.0 (24.6–91.4) P90 153.5; >6 h 511 (1.16%)
Age: 40-64	n = 16,578 53.0 (23.0–100.0) P90 165.0; >6 h 244 (1.47%)	n = 27,139 52.9 (26.3–96.1) P90 159.2; >6 h 290 (1.07%)
Age: 65+	n = 4,094 59.0 (25.0–106.0) P90 166.0; >6 h 48 (1.17%)	n = 6,627 60.1 (29.1–103.8) P90 166.1; >6 h 55 (0.83%)
Sex: Female	n = 29,718 50.0 (22.0–94.0) P90 157.0; >6 h 453 (1.52%)	n = 49,509 51.4 (25.6–92.0) P90 152.2; >6 h 500 (1.01%)
Sex: Male	n = 42,871 50.0 (22.0–95.0) P90 161.0; >6 h 612 (1.43%)	n = 69,399 51.0 (25.3–93.1) P90 156.9; >6 h 778 (1.12%)
Acuity: Resuscitation	n = 638 17.5 (5.0–60.0) P90 133.3; >6 h 18 (2.82%)	n = 397 23.6 (12.1–72.8) P90 162.3; >6 h 5 (1.26%)
Acuity: Emergency	n = 3,752 19.0 (9.0–49.0) P90 148.0; >6 h 96 (2.56%)	n = 1,864 34.3 (13.8–86.1) P90 191.5; >6 h 64 (3.43%)
Acuity: Urgent	n = 9,705 52.0 (20.0–103.0) P90 183.0; >6 h 300 (3.09%)	n = 8,840 52.0 (23.0–95.8) P90 162.2; >6 h 155 (1.75%)
Acuity: Less Urgent	n = 46,412 54.0 (26.0–96.0) P90 153.0; >6 h 419 (0.90%)	n = 84,705 49.8 (25.8–87.3) P90 142.2; >6 h 683 (0.81%)
Acuity: Non Urgent	n = 12,082 46.0 (17.0–92.0) P90 167.0; >6 h 232 (1.92%)	n = 23,102 59.8 (27.1–116.9) P90 187.2; >6 h 371 (1.61%)

Table S7: Fixed test-population sensitivities

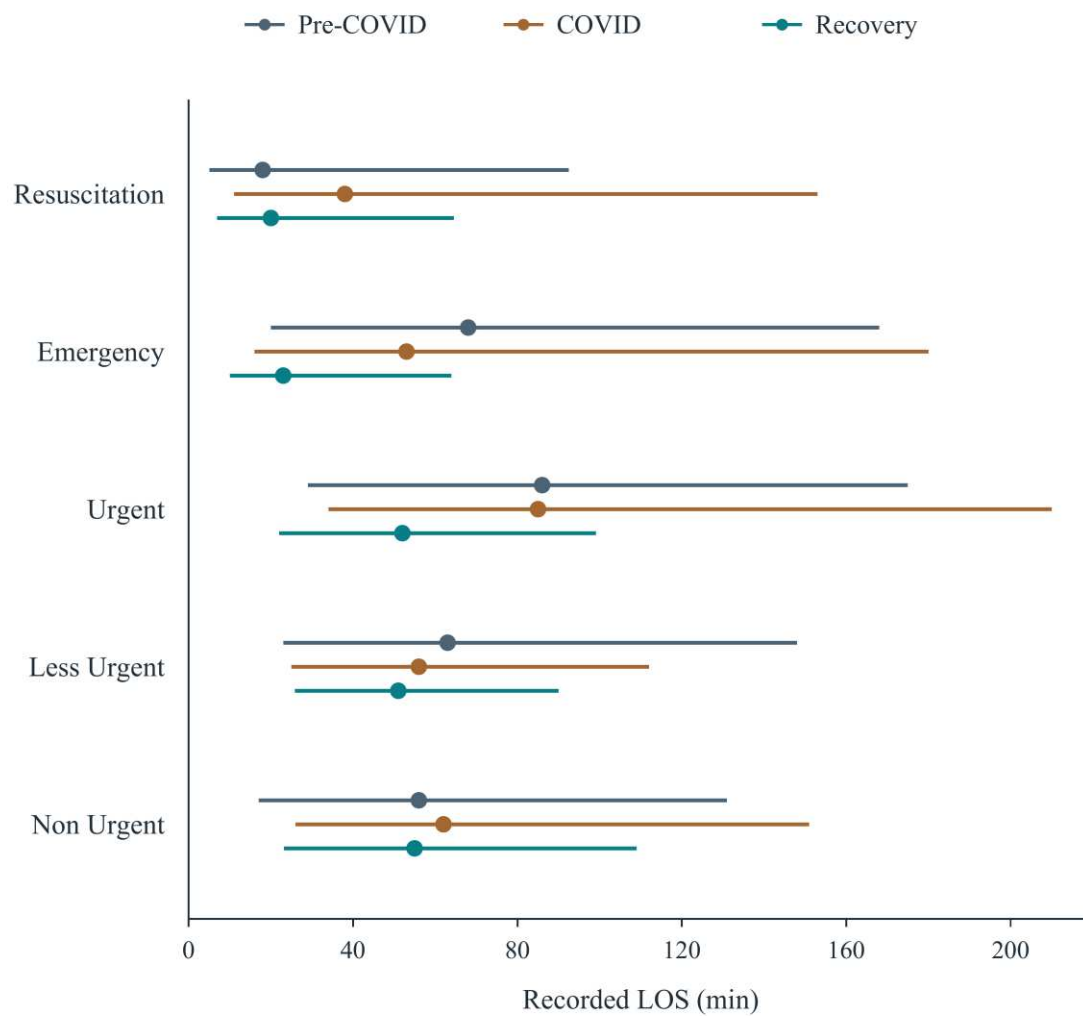
Sensitivity	Encounters / patients	M0 MAE / coverage (%)	M0R MAE / coverage (%)	M1 MAE / coverage (%)	M2 MAE / coverage (%)
Absent from development/calibration arrival windows	63,049 / 27,564	49.49 / 91.5	48.51 / 87.7	47.65 / 91.9	47.98 / 91.6
Absent from all earlier scoped arrivals, including gaps	60,553 / 26,846	49.55 / 91.4	48.59 / 87.7	47.71 / 91.8	48.03 / 91.6
Omit 2025 Q4	167,153 / 51,414	50.00 / 91.1	49.09 / 87.5	48.28 / 91.3	48.61 / 91.1
Include positive Error-coded encounters	191,506 / 55,632	49.25 / 91.5	48.30 / 87.8	47.66 / 91.6	48.02 / 91.4

Table S8: Training-tail refits on the unchanged test cohort

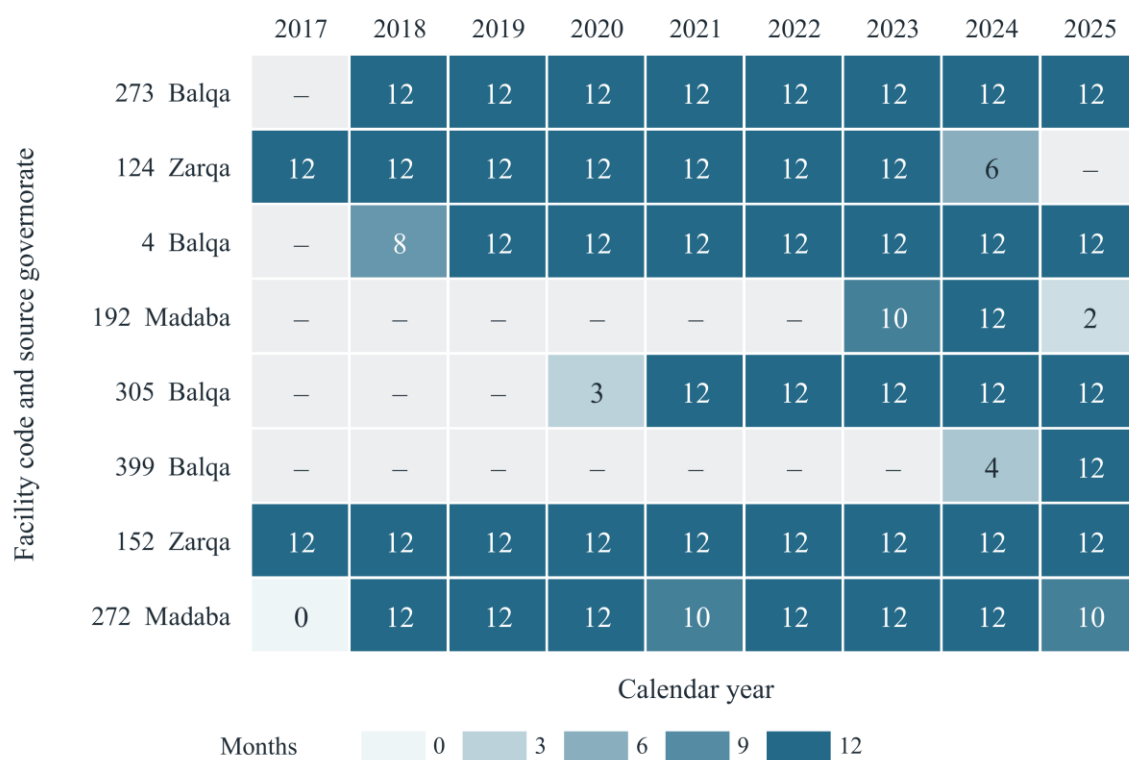
Model	Training encounters / patients	Test MAE / MedAE / RMSE (min)	MAE change from parent (min)	Coverage (%) / mean width (min)
M1, ≤72 h training	168,751 / 54,772	47.65 / 33.08 / 78.89	-0.01111	91.6 / 311.68
M2, ≤72 h training	168,751 / 54,772	48.02 / 33.53 / 79.09	0.00485	91.4 / 306.88



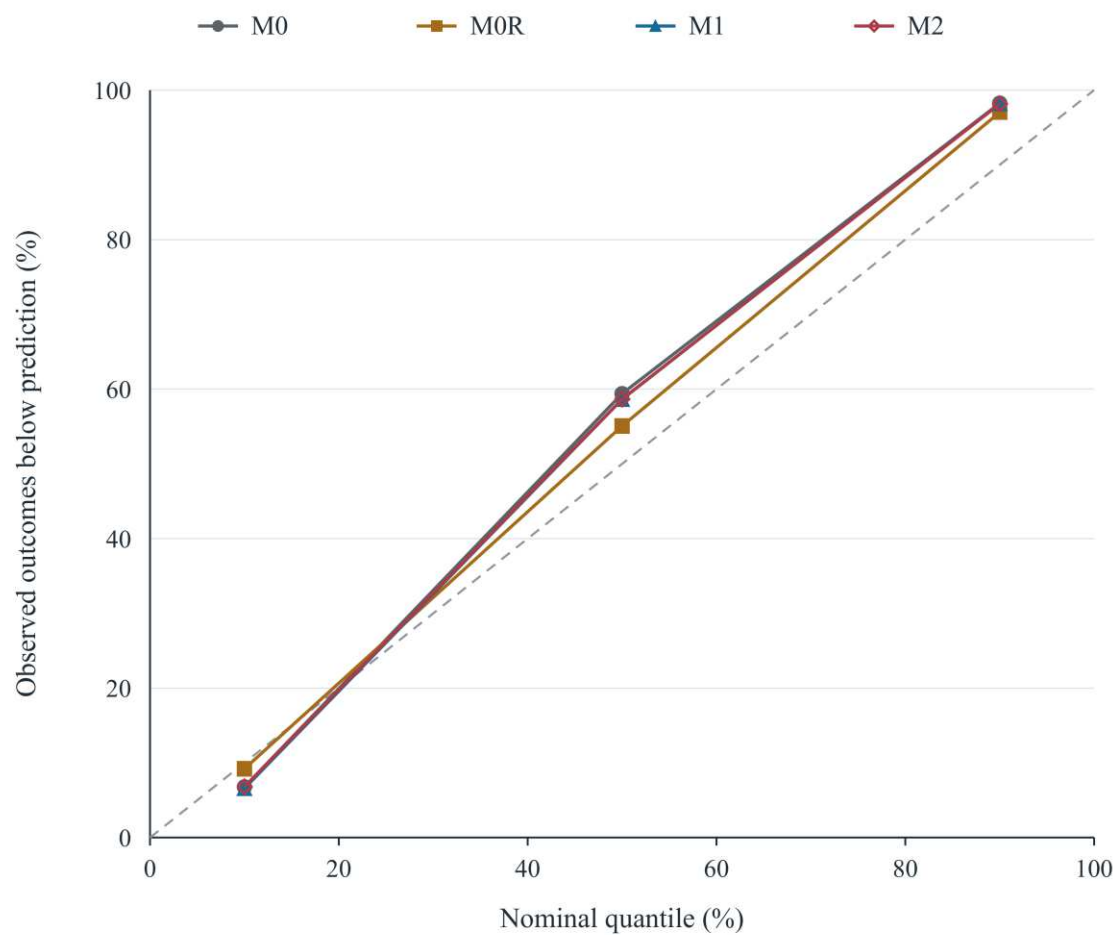
Supplementary Figure S1: Age and sex distributions of recorded stay across the three phases



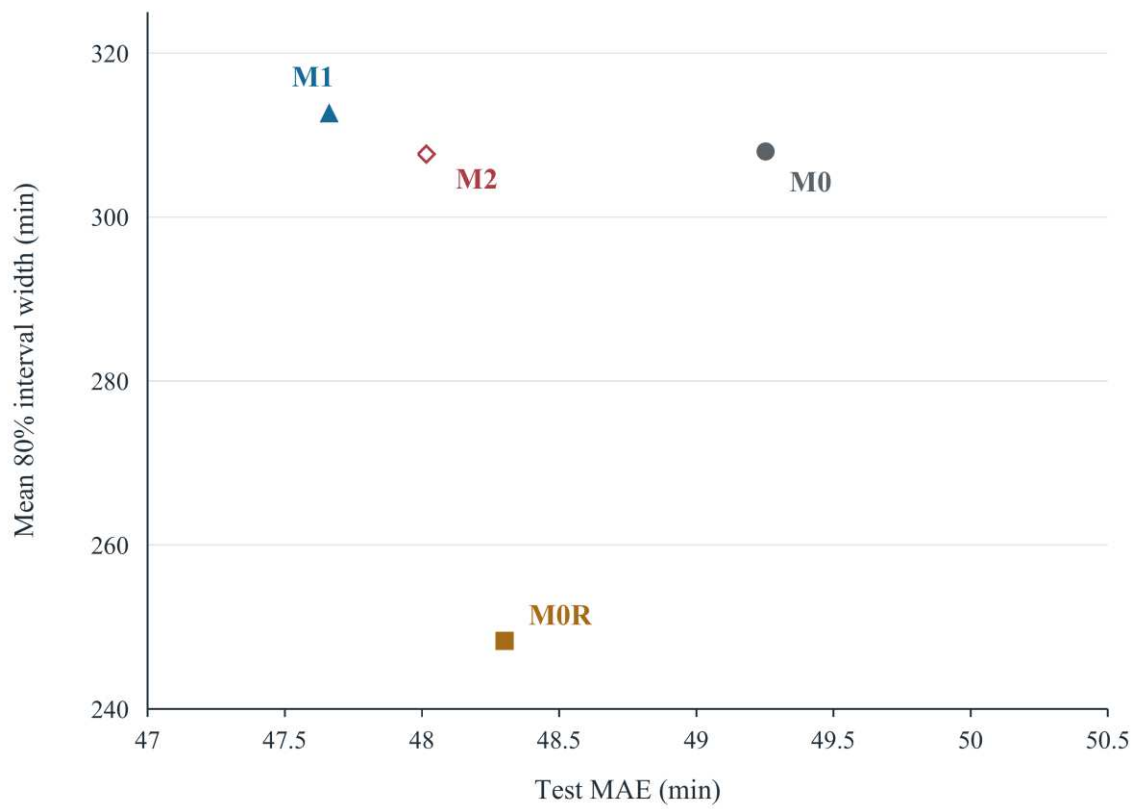
Supplementary Figure S2: Recorded stay within the source acuity categories



Supplementary Figure S3: Available months within the selected facility windows



Supplementary Figure S4: Observed fractions below fitted quantiles in the test cohort



Supplementary Figure S5: Point error and prediction interval width in the same later encounter cohort

All four models use the same 191,497 test encounters. MAE = mean absolute error; M0 = historical benchmark; M0R = recent 90-day benchmark; M1 = quantile regression; M2 = LightGBM.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [GITHUBREPOSITORYVIDEO1.mp4](#)
- [LOGSCHEMAMODELSCOMPARISONVIDEO2.mp4](#)
- [NOTEBOOKVIDEOVIDEO3.mp4](#)