Natural Resources for Human Health
eISSN: 2583-1194

Original Research

Received 14 Aug 2026
Revised 14 Sep 2026
Accepted 02 Oct 2026

Natural Resources for Human Health 2026, Vol. 6 (23s): 183–197

KEYWORDS:
clinical deterioration; early warning systems; artificial intelligence; external validation; calibration; health equity.

External Validation, Calibration, and Equity of Artificial Intelligence Early Warning Models for Clinical Deterioration in Hospitalized Patients: A Systematic Review and Exploratory Meta-Analysis

Ibrahim Daoud[1], Gawahir Murad Abdelrahman Alzein[2], Suliman Mahmoud[3], Khansaa Mohammed Ahmed Albadawi Fadlalmawla[4], Ahazeej Gurashi[5], Dalia Omar[6], Fatima Habib Allah Salama[7], Maysaa Ahmed[8], Alaa Hussien Abalmajed Jeaballa[9], Ghaida Hussien AbdAlmajed Jeaballa[10], Aiyda Mohammed Mohammed Ahmed[11], Malaz Taglwy[12], Aseel Alswairky[13], Fath Elrahman Elrasheed[14*]

[1]Department of Obstetrics and Gynecology, Alneelain University; Batterjee Medical College, Aseer Campus, Abha, Saudi Arabia. Email: Ibrahimdaaoud2020@gmail.com
[2]Obstetrics and Gynecology Consultant, Maternity and Children Hospital Alkharj, Alkharj, Saudi Arabia. Email: totaelzein2017@gmail.com
[3]Registrar Of Family Medicine, Armed Forces Hospital Southern Region, Ministry of Defense Health Services, Khamis Mushait, Saudi Arabia Email: Sulimanmahmoud29103@gmail.com; ORCID: 0009-0003-4103-7731
[4]Obstetrics and Gynecology Consultant, Almaali Hospital Hafralbatin, KSA Email: khanosobg@gmail.com
[5]Specialist in Obstetrics and Gynecology, Haya National Hospital, Khamis Mushait, Saudi Arabia. Email: Ahazeej.g@hotmail.com; ORCID: https://orcid.org/0009-0009-9667-5291
[6]Consultant of Obstetrics and Gynecology, Alsaedy Hospital, Al Jumom, Makkah Al mukarramah, Saudi Arabia. Email: Drdaliaomar18@gmail.com
[7]Consultant of Obstetrics and Gynecology, Adele Medical Clinics Complex, Riyadh, Saudi Arabia. Email: fatima19801111@gmail.com
[8]Obstetrics and Gynecology Consultant, Dallah Hospital, Riyadh, Saudi Arabia Email: maysaaahmed00@gmail.com
[9]Internal Medicine Consultant, Alrass General Hospital, Qassim Cluster, Saudi Arabia. Email: alaahussien2018@outlook.com
10Internal medicine consultant, Sumiara General Hospital Hail Cluster, Saudi Arabia. Email: drghaida87@gmail.com
[11]Home Health care Department, Family Medicine Registrar, King Khalid Hospital, Najran Email: draiydamohd@gmail. com; Tel: 00966567067400
[12]Consultant of Obstetrics and Gynecology, Wellness Experts Medical Center, Riyadh, Saudi Arabia. malazbabekir990@gmail.com
[13]Department of Obstetrics and Gynecology, Specialized Medical Center (SMC1), Riyadh, Saudi Arabia. Email: alswairkyaseel@gmail.com
[14]Assistant professor of Obstetrics and Gynaecology, Faculty of Medicine, Najran University – Saudi Arabia. Email: fathsaed@yahoo.com; ORCID: https://orcid.org/0009-0001-9764-5256
Corresponding author: Fath Elrahman Elrasheed; fathsaed@yahoo.com

Natural Resources for Human Health
eISSN: 2583-1194

ABSTRACT:

Background: Evaluation of a deterioration model in a new institution is not the same as internal validation and discrimination alone does not establish reliable risk estimates, fair performance or better outcomes. These differences are particularly relevant when the development provenance of proprietary scores is not disclosed.

Objective: To describe geographic external-validation evidence for data-derived inpatient deterioration models; calibrate and sociodemographic equity reporting; and perform narrowly defined exploratory synthesis when compatible model-specific data were available.

Methods: Europe PMC was searched until 3 September 2026 with two fully paginated queries; one AI assistant screened, assessed reports and extracted data; provisional PROBAST+AI-informed domain appraisal without independent duplicate review was performed only once. Eligible primary reports tested a fixed model in geographically separate institutions among adult general acute-care inpatients initially free of target event. Commercial models with unverifiable training site overlap and dedicated equity/implementation reports were synthesized separately. AUROCs from seven external M-CURES evaluation cohorts in one report were synthesized on the logit scale using restricted maximum likelihood and modified Hartung–Knapp intervals.

Results: The search returned 1,765 records; after removing 44 identifier duplicates there were 1,721 lefts. After title and abstract screening, we sought 110 reports; 27 weren't obtained and 83 were assessed. Thirty-four reports remained (18 geographic-validation confirmed reports, 13 provenance-unverified commercial score evaluations, three equity/implementation context reports). These are report counts not independent studies or unique populations counts. No overall total participant count was calculated because reports overlapped with institutional cohorts. For M-CURES seven external cohorts had 8335 admissions and 1304 qualifying events across 12 medical centers. Exploratory pooled AUROC was 0.811 (95% CI 0.787–0.834), I^2 45.7%, 95% prediction interval 0.757–0.856. Calibration and equity results were not sufficiently standardized for quantitative synthesis.

Conclusions: The retrieved evidence supports model-specific assessment not overall "AI early-warning accuracy." Geographic validation did not establish calibration, fairness or clinical effectiveness; the M-CURES synthesis only reports site-grouped performance in one coordinated report and should not be read as seven independent replications nor generalized to other deterioration models; single index coverage, unavailable reports and single assistant assessment substantially reduce confidence and require human verification before clinical/publication use.

1. INTRODUCTION

Inpatient deterioration models combine repeatedly collected clinical data to predict patients at risk of cardiac arrest, death or unplanned transfer to intensive care. Eligible literature includes gradient-boosted ensembles, deep neural networks, penalized models and proprietary data-derived indices with materially different predictors, prediction times and target events (e.g., eCARTv5 targets overlapping but nonidentical manifestations of deterioration; Mayo Clinic Early Warning Score and DeepCARS target overlapping but nonidentical manifestations). [1, 2, 3]

An assessment in an unused institution tests geographic transportability; a random patient split or later sample from the same hospitals does not. Calibration is agreement between predicted and observed risk; discrimination is separation of patients or observations with and without outcome. These are different performance dimensions that should be distinguished when synthesizing model evaluations. [4]

Equity asks additional questions: do performance differ across sociodemographic groups; do users access the model similarly; do clinical benefits differ after deployment? These cannot all be answered simultaneously (the CONCERN equity evaluation and the CHARTwatch sociodemographic evaluation show different levels of assessment). [5, 6]

Objectives and review question. The first objective was to describe confirmed geographic external validation of data-derived deterioration models in adult acute-care inpatients; second objectives were calibration, sociodemographic equity, applicability and clinical utility reporting; proprietary score provenance uncertainty; whether any compatible model specific numerical synthesis was possible.

The review question was formulated using population, index model, comparator, outcomes, timing and setting. Population was adult-service acute inpatients initially free of the target event; index models were fixed data derived deterioration scores; comparators were optional (conventional warning scores) and outcomes were general clinical deterioration or a prespecified qualifying composite; timing and unit of analysis were extracted separately rather than assumed to be equivalent across models.

2. MATERIALS AND METHODS

Protocol, registration and reporting: a working execution record was kept locally but the review was not prospectively registered and no registry identifier is claimed. PRISMA 2020 and TRIPOD-SRMA guided reporting; use of these guidelines should not be taken as certification of methodological completeness. [7, 8]

A post hoc amendment to report assessment created a separate stream for unchanged commercial/proprietary scores whose training-site overlap could not be verified (the provenance opacity was not intended). A targeted recheck of excluded abstracts for named commercial comparators returned five additional reports for assessment; four were retained and one was excluded. Another source correction returned the 2018 DEWS report because its methods had a separate hospital evaluation even though the abstract said temporal split. All six were already in deduplicated index results so are not new search records. A contextual stream kept dedicated sociodemographic equity evaluations and parent CONCERN trial but not with predictive validation cohorts.

Eligibility criteria

Primary eligibility was an adult acute-care ward or general acute inpatient population, eligible data-derived deterioration model, qualifying outcome and evaluation of a fixed model in institution or geographically separate health system not used for model development; cohorts from adult care starting at age 16 (and cohorts initially free of target event) allowed with setting and applicability limitations noted; penalized regression ensembles and proprietary data-derived indices allowed but algorithm families labelled rather than all scores deep learning.

Pediatric-only, emergency-department-only, ICU-only, long term acute care sepsis-only respiratory support-only studies were excluded unless separately extractable eligible population and general deterioration evaluation were available. Ordinary clinical regression scores without an eligible ML/proprietary index model were excluded. Random split, temporal only and cross validation only development of new models did not meet the primary external validation definition. A locally recalibrated or reoptimized vendor model was not treated as unchanged external validation.

For the provenance-unverifiable stream a report had to assess a previously specified general deterioration score without fitting that score to data but could have no verifiable separation of training from hospital. Actual same cohort predictive performance had to be assessed: cross-study benchmark tables and association only odds ratios were not accepted as predictive validation. This stream does not establish geographic externality. High dependency, post ICU, COVID specific, surgical and trauma cohorts were not assumed to represent an unrestricted general medical ward population.

Information sources and search

The systematic bibliographic source was Europe PMC searched on 3 September 2026 via its public application programming interface. Europe PMC aggregates indexed records but this was one index search not MEDLINE, Embase, CINAHL, Scopus or Web of Science searches. Two queries combined deterioration/early-warning concepts; data-derived modeling terms; inpatient settings while a second named-model query was performed. Exact Boolean strings, timestamps, pagination files and retrieval totals are preserved in audit materials and Appendix A.

The queries had a publication date upper bound of 3 September 2026 and no query level language restriction. Assessment was on accessible English material (no formal multilingual translation process was used); web retrieval and source checking targeted the full texts of indexed candidates and methodological guidance; no additional off-index study was silently added to PRISMA denominator; authors were not contacted; subscription-only database or library access was not used.

Deduplication and study selection. Identifier-based deduplication used the retrieved bibliographic identifiers before screening. One AI assistant did title screening, abstract screening, full text assessment and adjudication; there were no two independent reviewers nor human consensus process. Persistent record identifiers linked the original search metadata (search terms, decision, reason etc.) source location (original publication). Preliminary classifications were updated if methods links changed eligibility or model provenance links changed eligibility or publication-version links changed eligibility.

Full text was sought for 110 reports (five of which were returned during the commercial-comparator recheck and one after a geographic validation source correction). Reports for which only abstract or bibliographic content were found were labeled "not retrieved" rather than considered fully assessed. Earlier reports superseded by a full publication were linked not counted as independent evidence; multiple eligible reports from a common cohort could remain in the narrative synthesis with overlap flagged explicitly.

Extraction and data handling

Extraction included model and algorithm, development and evaluation geography, cohort size (and qualifying events when available), outcome and prediction horizon, analysis unit, discrimination/uncertainty, calibration, sociodemographic analyses, clinical utility and overlap/applicability concerns. Main-text XML or targeted native full-text extracts were retained as evidence; an abstract was not substituted for inaccessible methods. Source excerpts and source URLs are given with extraction records.

Admission, patient and score-observation denominators were identified

A C-index was not renamed an AUROC; operating characteristics were not converted into AUROCs and absent confidence intervals were not invented. Baseline composition of subgroups was not equity evaluation. Calibration of a component score was not attributed to a combined model. Grand total of unique participants from possibly overlapping reports was not calculated.

Risk of bias and applicability

The assessment was a PROBAST+AI-informed provisional appraisal, not an independently adjudicated application of every signaling question. Participants and data sources; predictors; outcome; analysis notes were made for each and applicability separately (the updated tool differentiates evaluation risk of bias from development quality so an omission in reporting wasn't automatically high risk). [9]

Important issues were the case mixes selected, case-control sampling, predictors available at decision time, composite outcomes ascertained, repeated scores within admissions, missing data processing, data dependent thresholds, local adaptation and incomplete uncertainty reporting. No numerical "quality score", formal overall low-risk percentage or GRADE certainty rating was given. The domain-level evidence notes are provided for independent verification.

Synthesis and statistical analysis

The main synthesis was narrative and model specific. Different models, versions, outcome definitions, prediction horizons and patient versus observation units were not pooled into an overall AI estimate. A pragmatic minimum of three compatible external evaluation cohorts was used to consider numerical synthesis but was not taken as proof of clinical comparability. [4]

Only M-CURES provided an immediately usable set of geographically separated cohorts with a common fixed model, endpoint definition and uncertainty published for the selected exploratory analysis. Seven external evaluation cohorts came from one coordinated report; cohort G combined six medical centers. Internal Michigan validation was excluded. Outcome was deterioration within first five days after admission (death, invasive ventilation, heated high-flow nasal cannula or intravenous vasopressors). Scores were updated every four hours; this is not rolling five-day forecast at every score time. Discrimination was done at admission level by threshold crossing before endpoint whereas calibration was done per four-hour prediction window. [10]

Published Figure 1 AUROCs and confidence limits were visually compared with the archived figure. Standard errors were approximated as width of logit confidence limit divided by twice 1.959964. Random-effects synthesis was restricted maximum likelihood on the logit scale (modified Hartung–Knapp variance multiplier bounded below by one) and t distribution with six degrees of freedom; approximate prediction interval used five degrees of freedom plus between-cohort variance plus mean variance. Methods were chosen according to prediction-model meta-analysis guidance but their application here is exploratory. [4, 10]

Sensitivity analyses were omitted the combined six-center cohort and each cohort in turn. No publication-bias test, funnel-plot inference, meta-regression or pooled clinical-effect estimate was performed. Rounded published intervals, common model-development uncertainty, reporting practices and unknown cross-site patient linkage constrain the inferential interpretation. Computations were done in Python, NumPy and SciPy; version information, inputs and runnable code are provided.

3. RESULTS

Search and selection. The general query returned 1,619 records; the named-model query returned 146 so total retrieved records were 1765 (after removing 44 identifier-based duplicates). Of these 1301 were excluded at title assessment and 420 abstracts assessed; after abstract adjudication and post hoc comparator recheck 310 records were excluded at abstract stage and 110 reports sought.

Twenty-seven reports were not obtained as full text. Eighty-three reports were assessed (49 excluded and 34 retained across the three synthesis streams above), stream counts were 18 confirmed geographic-validation reports; 13 provenance-unverifiable commercial-score evaluations; 3 equity/implementation context reports. Figure 1 shows the reconciled flow and supplementary ledger gives every decision made.

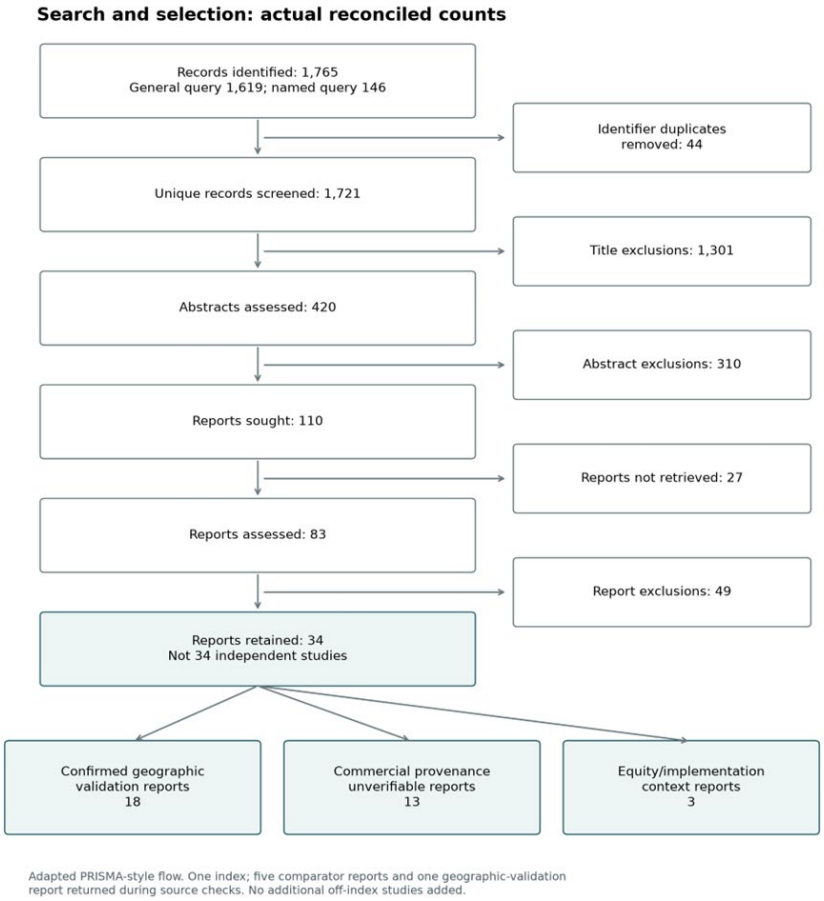


Figure 1. Reconciled public-index search and report selection. Counts refer to reports, not independent studies.

The report count cannot be taken to mean 34 independent investigations. Presbyterian Medical Center VC-MAES reports have the same setting and enrollment period; CONCERN equity analysis and parent trial apply to the same underlying trial; eCART reports may contain overlapping development or evaluation populations. The ledger also lists reports not obtained (e.g., earlier deep-learning warning-score validations) so missing full text isn't evidence of ineligibility but rather a coverage limitation. [1, 11, 12]

Confirmed geographic-validation evidence. Table 1 shows the 18 primary reports. Evaluations differ in clinical case mix, target event, horizon and measurement unit; table is not a ranking of models nor does it support indirect performance comparisons.

The 2018 DEWS report tests separately Hospital B with 1,772 patients (four cardiac-arrest patients only). The widely cited AUROC of 0.850 is for Hospital A's temporal test not the external-center numerical result; the external curves are shown graphically and no external confidence interval was extracted; the event-vector counts in the table are inconsistent (82 vs 88 in the narrative); these limitations are kept in the evidence table rather than replaced by the headline metric.15 [13]

Confirmed geographic-validation evidence

Table 1 summarizes the 18 primary reports. These evaluations differ in clinical case mix, target event, horizon, and measurement unit; the table is not a ranking of models and does not support indirect performance comparisons.

The 2018 DEWS report separately tests 1,772 patients at Hospital B, with only four cardiac-arrest patients. Its widely cited AUROC of 0.850 belongs to Hospital A's temporal test, not to the external-center numerical result; the external curves are shown graphically and no external confidence interval was extracted. The report also gives inconsistent event-vector counts for Hospital B, 82 in the table and 88 in the narrative. These limitations are retained in the evidence table rather than replaced by the headline metric. [13]

Table 1. Confirmed geographic-validation reports. Numbers are source-specific, not comparable model rankings

Report/model	Evaluation population and geography	Relevant result and interpretation
R0041: eCARTv5 (2025) [1]	Illinois development; external CT/RI/WI. Retrospective 1,769,461; prospective 205,946 admissions.	24-h observation AUROC 0.834 (0.834–0.835) retrospectively and 0.828 (0.827–0.829) prospectively. Observation-level limits; not pooled.
R0057: VC-MAES (2025) [14]	Severance development; prospective Keimyung Dongsan validation. 6,039 patients; 217 events.	24-h major adverse events: AUROC 0.918 (0.901–0.935).
R0099: VC-MAES (2025) [11]	Severance development; Presbyterian Medical Center. 24,359 admissions; overlaps R0136.	24-h major adverse events: observation AUROC 0.870 (0.855–0.885).
R0136: VC-MAES (2025) [12]	Presbyterian Medical Center; 30,785 general-ward admissions within a 32,720-admission mixed ward/ICU report. Overlaps R0099.	24-h major adverse events: reported AUROC 0.869 (0.854–0.884); exact discrimination denominator is not separately clear.
R0538: Multimodal DRNN (2025) [15]	University of Chicago development; University of Wisconsin validation. 248,055 encounters.	Death/ICU transfer at ≥24-h lead time: external AUROC 0.870 (0.866–0.874); no-lead-time AUROC 0.964. These are not the same task.

Report/model	Evaluation population and geography	Relevant result and interpretation
R0205: COVet (2024) [16]	VHA development; 37,011 non-VHA admissions at two US centers.	Death/ICU transfer within 24 h: non-VHA AUROC 0.86 (CI not reported). VHA internal/temporal results not substituted.
R0208: eCARTv5 comparative evaluation (2024) [17]	362,926 adult non-ICU ward encounters at seven Yale New Haven hospital sites; potential overlap with R0041.	Maximum pre-event encounter score: AUROC 0.895 (0.892–0.898); NEWS2 0.831; EDI 0.808. Not interchangeable with observation AUROC.
R0586: DeepCARS (2024) [18]	Severance general wards; 109,626 admissions; 55 cardiac arrests.	24-h cardiac arrest: AUROC 0.90 (0.89–0.90). Consecutive-trigger threshold analyses were data-dependent.
R0995: COEWS / LASSO model (2024) [20]	SEMI-COVID development; Argentina HIBA external cohort, n=906.	Respiratory failure/death: narrative AUROC 0.743 (0.703–0.784); points score 0.700 (0.654–0.745). Table 3 stratifies by vaccination; combined-cohort interpretation is not assumed.

Several reports show large scale transportability assessment but cohort size does not eliminate endpoint or unit of analysis differences. eCARTv5 report separates retrospective from prospective evaluations DeepCARS and VC-MAES reports are specific cardiac-arrest/ICU-transfer or major adverse event settings respectively COVID and post ICU models have narrower applicability [34–37]. [1, 3, 14, 19, 20, 21, 22, 23, 24, 35, 36, 37]

Commercial-model provenance uncertainty

Table 2 contains 13 reports evaluating previously specified commercial or proprietary deterioration scores without a verifiable training-hospital separation. Inclusion in this table is not an endorsement of model independence, lack of overlap in training, low risk of bias. If the paper also developed a new model only its unchanged commercial-score evaluation goes into this stream when we consider that model development. [25, 26, 27, 28, 29, 30, 31, 32, 33]

Table 2. Commercial/proprietary-score evaluations with unverified training-site separation. New locally trained models are not substituted for the commercial comparator.

Report/model	Evaluation population and geography	Relevant result and interpretation
R0003: Epic DI (2026) [25]	89,492 encounters in a US tertiary hospital; training-site overlap not verifiable.	12-h events: encounter AUROC 0.759 (0.755–0.762). No numeric threshold PPV was inferred from class balance.
R0048: Epic DI comparator (2025) [26]	New vital-sign model has 36,784 validation episodes; EDI comparator denominator is not separately established.	At sensitivity 0.73, EDI PPV is reported as 7.57% (Table 2 rounded 0.08). Do not attribute the new LightGBM AUROC to EDI.
R0406: Epic DI by service (2025) [27]	61,151 encounters: 50,897 medical and 10,254 surgical.	Last-score AUROC: medical 0.869 (0.856–0.878), surgical 0.677 (0.668–0.686). Last-score design is not prospective fixed-lead-time testing.

Report/model	Evaluation population and geography	Relevant result and interpretation
R0237: Visensia Stability Index (2024) [28]	Singapore high-dependency unit; 1,557 admissions; training institutions undisclosed.	Composite physiological/clinical deterioration: AUROC 0.61 (0.55–0.66). High-dependency case mix limits ward applicability.
R0814: CoMET (2024) [29]	Prospective University of Virginia evaluation, 5,184 visits; original development hospital not established in the retrieved report.	ICU-transfer AUC 0.733 (0.701–0.761); cardiovascular 0.725, cardiorespiratory 0.737. Calibration plots and sex-stratified results reported.
R0857: Original Epic DI threshold comparator (2024) [30]	Four Mayo sites; balanced sample 6,446 event and 6,446 control patients; overlaps R0882.	Original DI≥60 rule F1 0.48 (0.48–0.49). AUROC 0.985 belongs to the newly fitted contrastive model, not original DI.
R0882: Original Epic DI threshold comparator (2023) [31]	Four Mayo sites; 12,892 balanced event/control patients; overlaps R0857.	Any DI>60 rule: sensitivity 0.25, specificity 0.89; binary-rule AUC 0.57 (0.57–0.58). Not continuous-score AUROC; PPV not transportable from balanced sampling.
R0402: Epic DI in trauma (2022) [32]	Single-center adult trauma cohort, 1,325 patients; commercial derivation overlap unverified.	Inpatient mortality: maximum-DI AUROC 0.982 (0.959–0.996). This is not a 12/24-h general-ward composite estimate.
R1197: Epic DI with image-model comparison (2022) [33]	Single-center COVID sample; 500-patient heldout cohort.	EDI-alone AUROC 0.70 (0.64–0.74). The CXR-augmented local model is not the eligible unchanged commercial score.
R1444: Epic DI, COVID (2021) [34]	Michigan Medicine; 392 admissions, 103 qualifying outcomes.	AUROC 0.79 (0.74–0.84). Threshold 68.8: PPV 74%, sensitivity 39%; threshold explored in this cohort.

These studies are helpful for characterizing deployment performance and reporting gaps but cannot be pooled with confirmed geographic validations just because the investigators describe their work as a validation. The Epic DI medical/surgical evaluation, for example reports materially different discrimination by clinical service; the prospective CoMET report has calibration and sex-stratified results but retrieved report does not establish the institution of the original development cohort. Calibration and equity. The evidence records separate actual calibration assessment from ranking performance, event rates within risk groups and calibration of only one score component. Calibration was not pooled quantitatively because metrics, horizons, units and availability varied. In M-CURES the original Figure 1 has expected external calibration errors of 0.01–0.04 but in the narrative it says range is 0.02–0.04; we keep this extraction discrepancy rather than silently normalize for it. Score-window calibration and five-bin summary should not be interpreted as proof that individual risk is calibrated in all settings. The post ICU RSI evaluation shows another attribution problem: calibration reported for RS1 does not imply calibration of combined RSI; prospective CoMET report has predicted versus observed risk plots and sex stratified discrimination but still in provenance-unverifiable stream. The dedicated contextual equity and implementation reports are summarized in Table 3. Groupwise AUROC differences, threshold-specific differences, subgroup calibration, clinician use and treatment-effect heterogeneity all address different questions; no single one was treated as a complete fairness assessment. [27, 29]

Table 3. Dedicated equity and implementation-context reports; no independence from a parent trial is implied.

Report	Population	Question and principal finding
R0296: CONCERN equity [5]	Parent four-hospital trial; 60,893 encounters including acute and intensive care units.	No statistically significant demographic differences in effectiveness were reported; male use exceeded female use by 0.09 percentage points. Use and clinical-effect equity are not predictive-calibration equity; no equivalence claim.
R0301: CHARTwatch equity [6]	Toronto GIM, 12,877 hospitalizations; 2,552 very-high-risk flags; 634 death/ICU events.	Specificity: age<60 93%, age60–80 90%, age>80 84%; demographic differences require uncertainty and clinical-context interpretation.
R0302: CONCERN parent trial [38]	Same four-hospital trial; 60,893 encounters; mixed acute-care and ICU setting.	In-hospital mortality adjusted HR 0.644 (0.532–0.778); unanticipated ICU transfer HR 1.249 (1.093–1.426). Implementation effect, not independent predictive validation.

M-CURES also reports performance on age sex race ethnicity subgroups but precision is small and institution level heterogeneity makes interpretation difficult. CHARTwatch's specificity reported across age groups varies; this is a measurable operating characteristic difference not by itself proof of unjust treatment or unequal clinical benefit. CONCERN measures usage effectiveness so its parent trial cannot be counted as an independent replication of the equity analysis. [5, 6, 10, 38]

Exploratory M-CURES synthesis. Seven external evaluation cohorts had 8,335 admissions and 1,304 qualifying events from 12 medical centers. Cohort G combined six smaller centers and the source report did not link across institutions to generate a unique-person denominator. Table 4 gives the complete input (the seven cohorts came from one report, not seven independently conducted studies). [10]

Table 4. Published M-CURES Figure 1 inputs; all seven external cohorts are from Kamran et al., 2022. Cohort G combines six centers. [10]

Cohort	Admissions	Events	AUROC (95% CI)
A	2320	311	0.82 (0.79–0.84)
B	1320	249	0.83 (0.80–0.86)
C	1256	206	0.82 (0.78–0.85)
D	1073	155	0.79 (0.74–0.83)
E	965	136	0.84 (0.81–0.87)
F	794	155	0.77 (0.73–0.82)
G	607	92	0.78 (0.73–0.83)

The random-effects pooled AUROC was 0.811 (95% CI 0.787–0.834), with I^2 45.7%, between-cohort variance on the logit scale 0.01173, and a 95% prediction interval of 0.757–0.856. Cochran's Q was 11.05 with six degrees of freedom (P=0.087). These values were calculated for this review from the published rounded Figure 1 inputs, rather than reported as a new analysis by the original authors. [10]

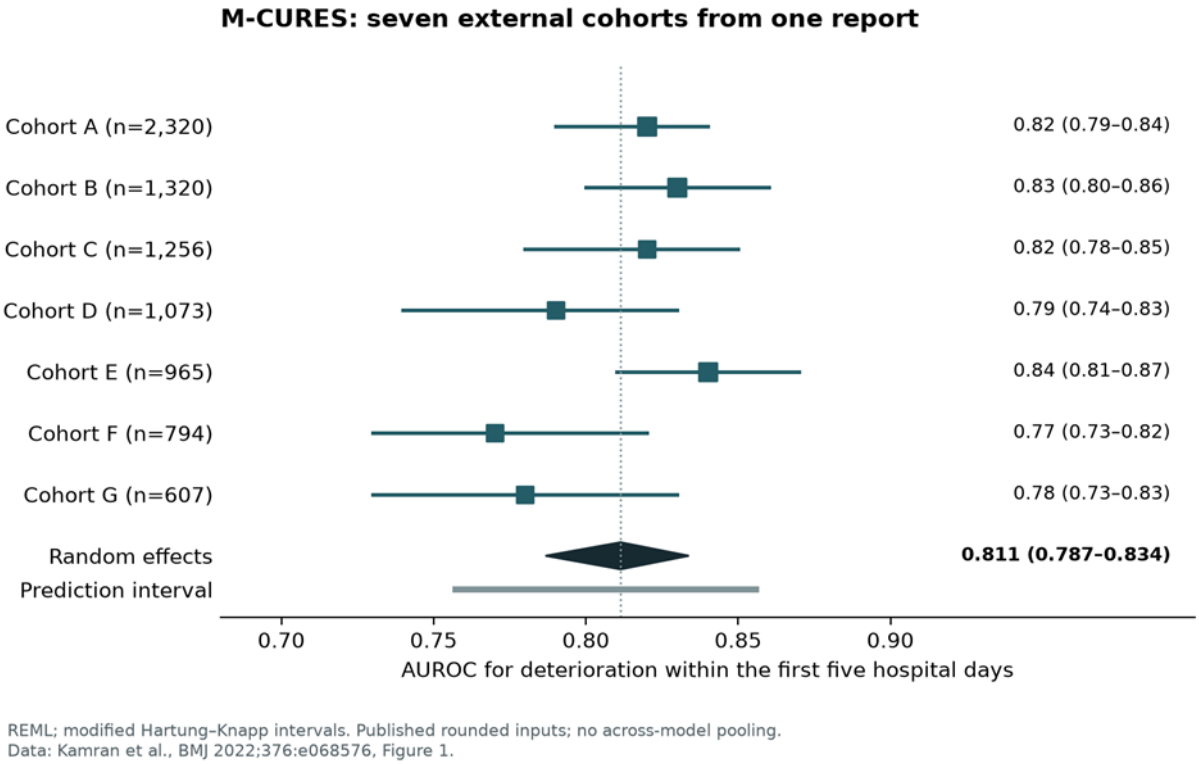


Figure 2. Exploratory M-CURES within-report external-cohort synthesis. Seven evaluation cohorts from one report; cohort G combines six centers.

Excluding combined cohort G yielded an AUROC of 0.815 (95% CI 0.789–0.839). Leave-one-cohort-out estimates are provided in the statistical supplement; these evaluate sensitivity to this site grouped inputs not independent study replication. No numerical synthesis was done for calibration, equity, clinical benefit or performance across AI models. Appraisal and applicability. The provisional domain notes describe limitations that matter to interpretation (not a summary quality score). Some COVID, post-ICU, high dependency and trauma populations limit generalizability. Repeated risk observations, heterogeneous composite events, thresholds chosen from study data and incomplete information on missingness or clustering complicate analysis. Proprietary training provenance creates another uncertainty about externality (not just reporting inconvenience). The assessment should not be interpreted as a formal completed PROBAST+AI evaluation by trained independent reviewers. No proportion of "low risk studies" is claimed. Main text nonreporting and incomplete access to supplements are distinguished from evidence that an assessment was not performed. Full report level notes accompany the manuscript so source checking and human adjudication can take place. [10]

4. DISCUSSION

Main findings. This review found heterogeneous retrieved evidence base (not one generalizable estimate of AI early-warning performance). Its clearest quantitative result is the M-CURES within-report cohort synthesis; all other evidence should be interpreted by model, version, outcome, population and prediction timing. Keeping a provenance-unverifiable stream means clinical evaluations of opaque vendor scores cannot be mislabeled as confirmed geographic validation. [4, 10]

Three distinctions are critical to interpretation. First, externality by geography is different from time validation or model implementation. Second discrimination doesn't establish calibration or consequences of acting on a threshold. Third subgroup

predictive performance, clinician use and clinical benefit are different measures of equity. The source-linked extraction in the review is organized around these distinctions not every reported metric as evidence of trustworthiness. [4, 5, 6]

Implications for practice. The results suggest cautious local evaluation (not scoring by largest reported AUROC) before deployment: an institution should first check that the model/version being evaluated and its inputs match its intended workflow; missingness and predictor availability; discrimination and calibration at intended horizon; threshold workload and clinically important missed events. If recalibration or modification is done, this change should be documented and the adapted model evaluated separately. [4, 7, 8, 9]

Equity evaluation should look at clinically meaningful operating characteristics and calibration of relevant groups; uncertainty and denominators; access, escalation and outcomes after implementation. A nonsignificant subgroup comparison does not mean equivalence. These are methodological implications of the review not that any included model is safe or unsafe for a particular hospital. [5, 6, 38]

Implications for research. Evaluations in the future would be more informative if they reported training institutions and dates, model versions (locked preprocessing and thresholds), source code/implementation specifications (where possible) and unit used to estimate uncertainty; common reporting of calibration intercepts/slopes or plots describing them, subgroup denominators and confidence intervals and alert workload would facilitate comparison. Clinically compatible independent external validation is needed before a model specific pooled estimate can be generalised beyond one coordinated report. [4, 8, 9]

Trial and observational deployment studies should link predictions to workflow and patient outcomes but not conflate outcome improvement with predictive accuracy. Equity investigations should be targeted for their question and not treat demographic composition or subgroup AUROC alone as a complete fairness measure. Prospective independent human verification of eligibility, extraction and risk-of-bias judgments is a necessary next step for this manuscript. [5, 6, 38]

Limitations. The systematic search was restricted to one public index, not a set of subscription bibliographic databases; the query and named-model list may miss eligible reports even if all returned pages are retrieved; 27 candidate reports were not obtained as full text (including potentially useful warning model evaluations); access failure can bias evidence base; no author contact, multilingual review process or formal search peer review was performed.

All screening, extraction and appraisal were done by one AI assistant. Classification error, incomplete extraction and wrong source interpretation are all risks (despite persistent identifiers and auditable ledger). The post hoc commercial-score amendment and contextual equity stream expanded the original scope and cannot be hidden from view. Main text availability varied between archived main text and targeted full-text extracts; inaccessible supplementary detail may contain additional information.

No prospective registration, independent duplicate assessment, full PROBAST+AI signaling-question adjudication or formal certainty grading were completed. Multiple reports shared cohorts and no unique patient denominator could be derived. Clinical impact and mixed setting equity reports were contextual evidence rather than validation evidence.

We used seven cohorts from one report (six centers combined into one cohort). Published AUROC and intervals were rounded; standard errors reconstructed; cross-site linkage and shared development uncertainty not modelled. The prediction interval is approximate with few cohorts and is not an estimate of performance in an arbitrary future hospital, disease or implementation. These restrictions preclude any overall AI-performance claim or inference that pooled AUROC establishes clinical benefit or fairness.

5. CONCLUSION

The documented search returned 18 geographic-validation reports, 13 provenance-unverifiable commercial score evaluations and three equity/implementation context reports. This is a report inventory with access limitations not 34 independent replications; the exploratory M-CURES synthesis gave an AUROC of 0.811 (95% CI 0.787–0.834) across seven external cohorts in one report only; no pooled estimate across models is justified for deployment decisions. Local validation, calibration, clinically grounded equity assessment and workflow and patient outcomes evaluation are needed before deployment decisions can be made. Independent human review and a broader bibliographic update must occur before this manuscript is used as definitive publication or clinical evidence source.

AUTHOR CONTRIBUTIONS

Ibrahim Daoud: Conceptualization, Methodology, Investigation, and Writing – Original Draft Preparation.

Gawahir Murad Abdelrahman Alzein: Investigation, Literature Search, Data Curation, and Writing – Review & Editing.

Suliman Mahmoud: Methodology, Validation, Formal Analysis, and Writing – Review & Editing.

Khansaa Mohammed Ahmed Albadawi Fadlalmawla: Investigation, Data Extraction, Data Curation, and Writing – Review & Editing.

Ahazeej Gurashi: Investigation, Literature Screening, Data Curation, and Writing – Review & Editing.

Dalia Omar: Methodology, Investigation, Validation, and Writing – Review & Editing.

Fatima Habib Allah Salama: Investigation, Data Extraction, and Writing – Review & Editing.

Maysaa Ahmed: Investigation, Data Curation, Visualization, and Writing – Review & Editing.

Alaa Hussien Abalmajed Jeaballa: Investigation, Literature Screening, and Writing – Review & Editing.

Ghaida Hussien AbdAlmajed Jeaballa: Investigation, Data Extraction, and Writing – Review & Editing.

Aiyda Mohammed Mohammed Ahmed: Investigation, Data Curation, and Writing – Review & Editing.

Malaz Taglwy :Investigation, Data Curation, Visualization, and Writing – Review & Editing.

Aseel Alswairky: Investigation, Validation, and Writing – Review & Editing.

Fath Elrahman Elrasheed (Corresponding Author): Conceptualization, Methodology, Supervision, Project Administration, Formal Analysis, Validation, Writing – Original Draft Preparation, and Writing – Review & Editing.

All authors contributed to the manuscript, critically reviewed its intellectual content, approved the final version for submission, and agreed to be accountable for their respective contributions to the work.

FUNDING

no funding

CONFLICT OF INTEREST

Authors has no conflicts of interest.

ETHICS AND DATA AVAILABILITY

The work used published reports and public bibliographic metadata, with no access to primary patient-level datasets. The authors should confirm any institutional requirements rather than treating this statement as a formal ethics exemption. Search logs, screening decisions, source-linked extraction, appraisal notes, numerical inputs, and runnable analysis code accompany the manuscript.

REFERENCES

[1] Churpek, M.M.; Carey, K.A.; Snyder, A.; Winslow, C.J.; Gilbert, E.; Shah, N.S.; Patterson, B.W.; Afshar, M.; Weiss, A.; Amin, D.N.; Rhodes, D.J.; Edelson, D.P. Multicenter Development and Prospective Validation of eCARTv5: A Gradient-Boosted Machine-Learning Early Warning Score. Crit. Care Explor. 2025, 7, e1232. https://doi.org/10.1097/cce.0000000000001232.

[2] Romero-Brufau, S.; Whitford, D.; Johnson, M.G.; Hickman, J.; Morlan, B.W.; Therneau, T.; Naessens, J.; Huddleston, J.M. Using machine learning to improve the accuracy of patient deterioration predictions: Mayo Clinic Early Warning Score (MC-EWS). J. Am. Med. Inform. Assoc. 2021, 28, 1207–1215. https://doi.org/10.1093/jamia/ocaa347.

[3] Cho, K.J.; Kim, J.S.; Lee, D.H.; Lee, S.M.; Song, M.J.; Lim, S.Y.; Cho, Y.J.; Jo, Y.H.; Shin, Y.; Lee, Y.J. Prospective, multicenter validation of the deep learning-based cardiac arrest risk management system for predicting in-hospital cardiac

arrest or unplanned intensive care unit transfer in patients admitted to general wards. Crit. Care 2023, 27, 346. https://doi.org/10.1186/s13054-023-04609-0.

[4] Debray, T.P.A.; Damen, J.A.A.G.; Snell, K.I.E.; Ensor, J.; Hooft, L.; Reitsma, J.B.; Riley, R.D.; Moons, K.G.M. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017, 356, i6460. https://doi.org/10.1136/bmj.i6460.

[5] Lee, R.Y.; Cato, K.D.; Dykes, P.C.; Lowenthal, G.; Jia, H.; Daramola, T.; Rossetti, S.C. Evaluating Equity in Usage and Effectiveness of the CONCERN Early Warning System. Appl. Clin. Inform. 2025, 16, 838–847. https://doi.org/10.1055/a-2630-4192.

[6] Colacci, M.; Pou-Prom, C.; Siddiqi, A.; Mamdani, M.; Verma, A.A. Evaluating sociodemographic bias in a deployed machine-learned patient deterioration model. JAMIA Open 2025, 8, ooaf158. https://doi.org/10.1093/jamiaopen/ooaf158.

[7] Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; Chou, R.; Glanville, J.; Grimshaw, J.M.; Hróbjartsson, A.; Lalu, M.M.; Li, T.; Loder, E.W.; Mayo-Wilson, E.; McDonald, S.; McGuinness, L.A.; Stewart, L.A.; Thomas, J.; Tricco, A.C.; Welch, V.A.; Whiting, P.; Moher, D. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021, 372, n71. https://doi.org/10.1136/bmj.n71.

[8] Snell, K.I.E.; Levis, B.; Damen, J.A.A.; Dhiman, P.; Debray, T.P.A.; Hooft, L.; Reitsma, J.B.; Moons, K.G.M.; Collins, G.S.; Riley, R.D. Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA). BMJ 2023, 381, e073538. https://doi.org/10.1136/bmj-2022-073538.

[9] Moons, K.G.M.; Damen, J.A.A.; Kaul, T.; Hooft, L.; Andaur Navarro, C.; Dhiman, P.; Beam, A.L.; Van Calster, B.; Celi, L.A.; Denaxas, S.; Denniston, A.K.; Ghassemi, M.; Heinze, G.; Kengne, A.P.; Maier-Hein, L.; Liu, X.; Logullo, P.; McCradden, M.D.; Liu, N.; Oakden-Rayner, L.; Singh, K.; Ting, D.S.; Wynants, L.; Yang, B.; Reitsma, J.B.; Riley, R.D.; Collins, G.S.; van Smeden, M. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ 2025, 388, e082505. https://doi.org/10.1136/bmj-2024-082505.

[10] Kamran, F.; Tang, S.; Otles, E.; McEvoy, D.S.; Saleh, S.N.; Gong, J.; Li, B.Y.; Dutta, S.; Liu, X.; Medford, R.J.; Valley, T.S.; West, L.R.; Singh, K.; Blumberg, S.; Donnelly, J.P.; Shenoy, E.S.; Ayanian, J.Z.; Nallamothu, B.K.; Sjoding, M.W.; Wiens, J. Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study. BMJ 2022, 376, e068576. https://doi.org/10.1136/bmj-2021-068576.

[11] Sim, T.; Cho, E.; Kim, J.; Kim, H.G.; Kim, S.J. Clinical Context Is More Important than Data Quantity to the Performance of an Artificial Intelligence-Based Early Warning System. J. Clin. Med. 2025, 14, 4444. https://doi.org/10.3390/jcm14134444.

[12] Kim, J.H.; Cho, E.Y.; Choi, Y.; Won, J.Y.; Cheon, S.H.; Kim, Y.A.; Lee, K.B.; Kim, K.J.; Kim, H.G.; Sim, T. Deep Learning-Based Early Warning Systems in Hospitalized Patients at Risk of Code Blue Events and Length of Stay: Retrospective Real-World Implementation Study. JMIR Med. Inform. 2025, 13, e72232. https://doi.org/10.2196/72232.

[13] Kwon, J.M.; Lee, Y.; Lee, Y.; Lee, S.; Park, J. An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest. J. Am. Heart Assoc. 2018, 7, e008678. https://doi.org/10.1161/jaha.118.008678.

[14] Sim, T.; Cho, E.Y.; Kim, J.H.; Lee, K.H.; Kim, K.J.; Hahn, S.; Ha, E.Y.; Yun, E.; Kim, I.C.; Park, S.H.; Cho, C.H.; Yu, G.I.; Ahn, B.E.; Jeong, Y.; Won, J.Y.; Cho, H.; Lee, K.B. Prospective external validation of a deep-learning-based early-warning system for major adverse events in general wards in South Korea. Acute Crit. Care 2025, 40, 197–208. https://doi.org/10.4266/acc.000525.

[15] Kotula, C.A.; Martin, J.; Carey, K.A.; Edelson, D.P.; Dligach, D.; Mayampurath, A.; Afshar, M.; Churpek, M.M. Comparison of Multimodal Deep Learning Approaches for Predicting Clinical Deterioration in Ward Patients: Observational Cohort Study. J. Med. Internet Res. 2025, 27, e75340. https://doi.org/10.2196/75340.

[16] Govindan, S.; Spicer, A.; Bearce, M.; Schaefer, R.S.; Uhl, A.; Alterovitz, G.; Kim, M.J.; Carey, K.A.; Shah, N.S.; Winslow, C.; Gilbert, E.; Stey, A.; Weiss, A.M.; Amin, D.; Karway, G.; Martin, J.; Edelson, D.P.; Churpek, M.M. Development and Validation of a Machine Learning COVID-19 Veteran (COVet) Deterioration Risk Score. Crit. Care Explor. 2024, 6, e1116. https://doi.org/10.1097/cce.0000000000001116.

[17] Edelson, D.P.; Churpek, M.M.; Carey, K.A.; Lin, Z.; Huang, C.; Siner, J.M.; Johnson, J.; Krumholz, H.M.; Rhodes, D.J. Early Warning Scores With and Without Artificial Intelligence. JAMA Netw. Open 2024, 7, e2438986. https://doi.org/10.1001/jamanetworkopen.2024.38986.

[18] Cho, K.J.; Kim, K.H.; Choi, J.; Yoo, D.; Kim, J. External Validation of Deep Learning-Based Cardiac Arrest Risk Management System for Predicting In-Hospital Cardiac Arrest in Patients Admitted to General Wards Based on Rapid Response System Operating and Nonoperating Periods: A Single-Center Study. Crit. Care Med. 2024, 52, e110–e120. https://doi.org/10.1097/ccm.0000000000006137.

[19] Pimentel, M.A.F.; Johnson, A.; Darbyshire, J.L.; Tarassenko, L.; Clifton, D.A.; Walden, A.; Rechner, I.; Watkinson, P.J.; Young, J.D. Development of an enhanced scoring system to predict ICU readmission or in-hospital death within 24 hours using routine patient data from two NHS Foundation Trusts. BMJ Open 2024, 14, e074604. https://doi.org/10.1136/bmjopen-2023-074604.

[20] Klén, R.; Huespe, I.A.; Gregalio, F.A.; Lalueza Blanco, A.L.; Pedrera Jimenez, M.; Garcia Barrio, N.; Valdez, P.R.; Mirofsky, M.A.; Boietti, B.; Gómez-Huelgas, R.; Casas-Rojo, J.M.; Antón-Santos, J.M.; Pollan, J.A.; Gómez-Varela, D. Development and validation of COEWS (COVID-19 Early Warning Score) for hospitalized COVID-19 with laboratory features: A multicontinental retrospective study. eLife 2023, 12, e85618. https://doi.org/10.7554/elife.85618.

[21] Cummings, B.C.; Blackmer, J.M.; Motyka, J.R.; Farzaneh, N.; Cao, L.; Bisco, E.L.; Glassbrook, J.D.; Roebuck, M.D.; Gillies, C.E.; Admon, A.J.; Medlin, R.P.; Singh, K.; Sjoding, M.W.; Ward, K.R.; Ansari, S. External Validation and Comparison of a General Ward Deterioration Index Between Diversely Different Health Systems. Crit. Care Med. 2023, 51, 775–786. https://doi.org/10.1097/ccm.0000000000005837.

[22] Pimentel, M.A.F.; Redfern, O.C.; Malycha, J.; Meredith, P.; Prytherch, D.; Briggs, J.; Young, J.D.; Clifton, D.A.; Tarassenko, L.; Watkinson, P.J. Detecting Deteriorating Patients in the Hospital: Development and Validation of a Novel Scoring System. Am. J. Respir. Crit. Care Med. 2021, 204, 44–52. https://doi.org/10.1164/rccm.202007-2700oc.

[23] Mauer, E.; Lee, J.; Choi, J.; Zhang, H.; Hoffman, K.L.; Easthausen, I.J.; Rajan, M.; Weiner, M.G.; Kaushal, R.; Safford, M.M.; Steel, P.A.D.; Banerjee, S. A predictive model of clinical deterioration among hospitalized COVID-19 patients by harnessing hospital course trajectories. J. Biomed. Inform. 2021, 118, 103794. https://doi.org/10.1016/j.jbi.2021.103794.

[24] Dziadzko, M.A.; Novotny, P.J.; Sloan, J.; Gajic, O.; Herasevich, V.; Mirhaji, P.; Wu, Y.; Gong, M.N. Multicenter derivation and validation of an early warning score for acute respiratory failure or death in the hospital. Crit. Care 2018, 22, 286. https://doi.org/10.1186/s13054-018-2194-7.

[25] Arvelo, I.C.; Shipley, K.; Wright, A.; Steitz, B.D. Machine learning on asynchronous clinical pages to predict clinical deterioration. JAMIA Open 2026, 9, ooag122. https://doi.org/10.1093/jamiaopen/ooag122.

[26] Romero-Brufau, S.; Smunyahirun, R.; Filhol, T.; Niederhauser, L.; Trakoolwilaiwan, T.; Singh, G. Vital Signs-Only Machine Learning Model for Acute Inpatient Deterioration: A Retrospective Multicenter Study. Mayo Clin. Proc. Innov. Qual. Outcomes 2025, 9, 100663. https://doi.org/10.1016/j.mayocpiqo.2025.100663.

[27] Caraballo, P.J.; Meehan, A.M.; Fischer, K.M.; Rahman, P.; Simon, G.J.; Melton, G.B.; Salehinejad, H.; Borah, B.J. Trustworthiness of a machine learning early warning model in medical and surgical inpatients. JAMIA Open 2025, 8, ooae156. https://doi.org/10.1093/jamiaopen/ooae156.

[28] Ehara, J.; Masuda, Y.; Hayashi, K.; Norisue, Y.; Fujitani, S. Efficacy of Early Warning Scores as the Prediction Tool for Detecting Patients With Acute Deterioration in a High Dependent Unit. Cureus 2024, 16, e71971. https://doi.org/10.7759/cureus.71971.

[29] Keim-Malpass, J.; Moorman, L.P.; Moorman, J.R.; Hamil, S.; Yousefvand, G.; Monfredi, O.J.; Ratcliffe, S.J.; Krahn, K.N.; Jones, M.K.; Clark, M.T.; Bourque, J.M. Prospective validation of clinical deterioration predictive models prior to intensive

care unit transfer among patients admitted to acute care cardiology wards. Physiol. Meas. 2024, 45, 065004. https://doi.org/10.1088/1361-6579/ad4e90.

[30] Salehinejad, H.; Meehan, A.M.; Caraballo, P.J.; Borah, B.J. Contrastive Transfer Learning for Prediction of Adverse Events in Hospitalized Patients. IEEE J. Transl. Eng. Health Med. 2024, 12, 215–224. https://doi.org/10.1109/jtehm.2023.3344035.

[31] Salehinejad, H.; Meehan, A.M.; Rahman, P.A.; Core, M.A.; Borah, B.J.; Caraballo, P.J. Novel machine learning model to improve performance of an early warning system in hospitalized patients: a retrospective multisite cross-validation study. EClinicalMedicine 2023, 66, 102312. https://doi.org/10.1016/j.eclinm.2023.102312.

[32] Mou, Z.; Godat, L.N.; El-Kareh, R.; Berndtson, A.E.; Doucet, J.J.; Costantini, T.W. Electronic health record machine learning model predicts trauma inpatient mortality in real time: A validation study. J. Trauma Acute Care Surg. 2022, 92, 74–80. https://doi.org/10.1097/ta.0000000000003431.

[33] Mu, E.; Jabbour, S.; Dalca, A.V.; Guttag, J.; Wiens, J.; Sjoding, M.W. Augmenting existing deterioration indices with chest radiographs to predict clinical deterioration. PLoS ONE 2022, 17, e0263922. https://doi.org/10.1371/journal.pone.0263922.

[34] Singh, K.; Valley, T.S.; Tang, S.; Li, B.Y.; Kamran, F.; Sjoding, M.W.; Wiens, J.; Otles, E.; Donnelly, J.P.; Wei, M.Y.; McBride, J.P.; Cao, J.; Penoza, C.; Ayanian, J.Z.; Nallamothu, B.K. Evaluating a Widely Implemented Proprietary Deterioration Index Model among Hospitalized Patients with COVID-19. Ann. Am. Thorac. Soc. 2021, 18, 1129–1137. https://doi.org/10.1513/annalsats.202006-698oc.

[35] Beals, J.; Barnes, J.J.; Durand, D.J.; Rimar, J.M.; Donohue, T.J.; Hoq, S.M.; Belk, K.W.; Amin, A.N.; Rothman, M.J. Stratifying Deterioration Risk by Acuity at Admission Offers Triage Insights for Coronavirus Disease 2019 Patients. Crit. Care Explor. 2021, 3, e0400. https://doi.org/10.1097/cce.0000000000000400.

[36] Cummings, B.C.; Ansari, S.; Motyka, J.R.; Wang, G.; Medlin, R.P.; Kronick, S.L.; Singh, K.; Park, P.K.; Napolitano, L.M.; Dickson, R.P.; Mathis, M.R.; Sjoding, M.W.; Admon, A.J.; Blank, R.; McSparron, J.I.; Ward, K.R.; Gillies, C.E. Predicting Intensive Care Transfers and Other Unforeseen Events: Analytic Model Validation Study and Comparison to Existing Methods. JMIR Med. Inform. 2021, 9, e25066. https://doi.org/10.2196/25066.

[37] Finlay, G.D.; Rothman, M.J.; Smith, R.A. Measuring the modified early warning score and the Rothman index: advantages of utilizing the electronic medical record in an early warning system. J. Hosp. Med. 2014, 9, 116–119. https://doi.org/10.1002/jhm.2132.

[38] Rossetti, S.C.; Dykes, P.C.; Knaplund, C.; Cho, S.; Withall, J.; Lowenthal, G.; Albers, D.; Lee, R.Y.; Jia, H.; Bakken, S.; Kang, M.J.; Chang, F.Y.; Zhou, L.; Bates, D.W.; Daramola, T.; Liu, F.; Schwartz-Dillard, J.; Tran, M.; Bokhari, S.M.A.; Thate, J.; Cato, K.D. Real-time surveillance system for patient deterioration: a pragmatic cluster-randomized controlled trial. Nat. Med. 2025, 31, 1895–1902. https://doi.org/10.1038/s41591-025-03609-7