



SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Research Paper

Open Access

Artificial Intelligence-Driven Clinical Decision Support Systems for Emergency Department Triage and Acute Care Prioritization: A Systematic Review of Accuracy, Safety, and Implementation Readiness

Ibrahim Daoud^{1*}, Osman Abdelazim Osman², Alaa Hussien Abalmajed Jeaballa³, Ghaida Hussien AbdAlmajed Jeaballa⁴, Ibtihal Abdalla Abdelmaged Abdalla⁵, Aseel Alswairky⁶, Mutaz M. Ahmed⁷, Abeer Salahaldin Mustafa⁸, Mohammed Abdelrazig Ahmed Elrofaie⁹, Suliman Mahmoud¹⁰

¹Department of Obstetrics and Gynecology, Alneelain University; Batterjee Medical College, Aseer Campus, Abha, Saudi Arabia. Ibrahimdaoud2020@gmail.com

²Neonatologist, Maternity and Children Hospital, Najran, Saudi Arabia. osman1995@hotmail.com

³Internal Medicine Consultant, Alrass General Hospital, Qassim Cluster, Saudi Arabia. alaaahussien2018@outlook.com

⁴Internal medicine consultant, Sumiara General Hospital Hail Cluster, Saudi Arabia. drghaida87@gmail.com

⁵Family medicine Registrar, Najran Health Cluster, Najran, Saudi Arabia. hola_kassala@yahoo.com

⁶Department of Obstetrics and Gynecology, Specialized Medical Center (SMC), Riyadh, Saudi Arabia. alswairkyaseel@gmail.com

⁸General Practitioner, Public Health Authority, Najran, Saudi Arabia. drmutaz2@yahoo.com

⁸Family Medicine Registrar, Najran Health Cluster, Saudi Arabia. Abeermust70@gmail.com

⁹Pediatric Consultant, Al-Shamly General Hospital Hail Cluster, Saudi Arabia. drrofaie87@gmail.com

¹⁰Registrar of Family Medicine, Armed Forces Hospital Southern Region, Ministry of Defense Health Services, Khamis Mushait, Saudi Arabia

Abstract

Background: AI is being increasingly investigated as a clinical decision support tool for ED triage and acute-care prioritization. Machine-learning, deep-learning, natural language processing (NLP) and large language model systems have shown good predictive performance but concerns remain about under-triage, over-triage, generalizability, workflow integration and implementation safety. We systematically reviewed the accuracy, safety and implementation readiness of AI-driven clinical decision-support systems for ED triage.

Methods: This review was conducted according to PRISMA 2020 recommendations. PubMed/MEDLINE, PubMed Central, Embase, Scopus, Web of Science and Cochrane Library were searched (citation chaining, reference-list screening and recent systematic reviews were also used). Of the 1076 records identified, 312 duplicates were removed and 764 records screened; 79 full-text reports were assessed and 65 excluded; fourteen studies were included in narrative synthesis. Meta-analysis was not performed because populations, AI architectures, triage systems, reference standards and outcome metrics are highly heterogeneous.

Results: Conventional machine learning, gradient boosting, neural networks, deep learning, natural language processing and large language models were included in the studies. AI systems generally outperformed conventional triage for critical illness prediction (hospitalization), intensive-care admission and mortality; reported AUROCs for early mortality were 0.962 and for 30-day mortality 0.925. Direct triage systems also had some success: KATE had 75.7% ESI assignment accuracy; prospective GPT-4 and GPT-4o studies had $\kappa = 0.899$ and multicenter $\kappa = 0.833$ with F1 = 0.897 agreement with expert triage but safety outcomes under-triage/over-triage subgroup performance and patient-important outcomes were inconsistently reported.

Conclusions: ED triage systems powered by AI have great potential as supervised clinical decision support tools but not yet enough evidence to justify autonomous deployment. Prospective multicenter validation, standardized safety metrics, subgroup analysis, workflow evaluation, governance and post-deployment monitoring are needed before widespread implementation.

Keywords: artificial intelligence; emergency department; triage; clinical decision support; machine learning; deep learning; large language models; patient safety; implementation readiness.

1. Introduction

EDs are busy places with high patient turnover, variable disease severity, time-sensitive decisions and frequent resource constraints. Triage is thus needed to triage patients who need urgent attention (i.e., access diagnostic/therapeutic resources), prioritize access to such resources and prevent avoidable delays in care. Conventional triage systems such as ESI, MTS, CTAS and KTAS provide standardised frameworks for urgency assignment but their performance is affected by clinician experience, incomplete information, workload and inter-rater variability so there are always risks of both under-triage and over-triage. AI has been increasingly investigated as a means to strengthen emergency triage and acute-care prioritization. Early CDSS based on structured variables (age, vital signs, comorbidities, arrival mode and conventional triage scores) mainly

predicted hospitalization, intensive-care admission, critical illness or mortality [1–5]. More recent work has incorporated NLP, free-text nursing assessments, chief complaints and deep learning architectures so that models can capture information not represented by traditional scoring systems [4–6]. This suggests AI could help stratify risk at the earliest point in emergency care pathway. Several studies have shown good predictive performance: Raita et al. improved prediction of critical-care needs and hospitalization with machine-learning models compared to ESI-based approaches [1]; Klug et al. strong discrimination for early mortality using gradient boosting [2]; Yu et al. machine-learning models based on initial nursing assessment could outperform conventional triage-based prediction [3]. Joseph et al. further showed that structured triage variables and textual chief-complaint information combined improved identification of critically ill patients [4]. Gains have also been reported for hospital admission and resource use prediction in pediatric and adult ED populations [5,8,10]. Clinical relevance of AI has moved beyond prediction of downstream outcomes to direct acuity classification as well. Ivanov et al. used machine learning to classify acute acuity directly; they found no agreement between doctors and machines [9]. developed KATE system which used ML and clinical NLP to support ESI assignment and had higher accuracy than routine nurse triage in validation sample [6]. Prospective evidence has also emerged. Liu et al. showed that mis-triage was reduced with machine learning assistance compared with traditional triage practice [9]. These studies are important because they evaluate AI in a way closer to actual triage workflow (rather than predicting future outcomes). Recently LLMs have been developed rapidly and new generation of clinician facing triage tools have appeared. GPT-4 and GPT-4o can read short clinical narratives and classify urgency without only having structured inputs available e.g. Paslı et al reported high agreement between GPT-4 and expert triage decisions in a prospective ED study [12], a multicenter study found GPT-4o strong agreement with expert reference standards across several emergency departments too [14]. Other studies show more variable performance and have emphasized accuracy depends on prompt design, case complexity, model version and reference standard used [13]. But predictive accuracy is not enough to prove clinical safety or implementation readiness: under-triage delays treatment of critically ill patients; over-triage crowds emergency departments and consumes scarce emergency resources. AI systems need therefore to be evaluated not only for discrimination and agreement but also class sensitivity (under-triage, over-triage), subgroup performance, calibration, robustness to incomplete data and effects on clinician behaviour. Implementation requires external validation, workflow integration, interpretability, governance, privacy protection, model-version control and monitoring for performance drift too. We conduct a systematic review on AI-driven clinical decision support systems for triage in emergency department and acute care prioritization. Accuracy, predictive performance, safety, comparison with conventional triage systems and human clinicians as well as implementation readiness are synthesised across machine learning, deep learning, NLP and LLM-based approaches. Special emphasis is put on distinguishing retrospective algorithmic performance from prospective clinical utility and on identifying conditions necessary for safe integration of AI into emergency-care workflows.

Methodology

2. 2.1 Study design and reporting framework

This systematic review examined AI-driven clinical decision support systems for ED triage and acute care prioritization (Table 1). The review was conducted and reported according to PRISMA 2020 statement [15]. Because substantial clinical and methodological heterogeneity was expected across study populations, triage systems, model architectures, outcomes and performance metrics we prespecified narrative synthesis as the main method of evidence synthesis. Statistical pooling was inappropriate where reporting; reporting was informed by SWiM guideline [16] where available. We reviewed AI systems used at triage or early ED assessment to classify urgency (critically ill), identify critically ill patients, predict deterioration, support disposition or prioritize acute-care resources. Technologies included conventional machine learning (ML) gradient boosting random forests neural networks deep learning natural-language processing LLMs.

2.2 Review question

The review question was structured using a Population–Intervention–Comparator–Outcome–Study Design framework. The population was adults and children presenting to emergency departments or ED-derived clinical records or realistic triage scenarios. The intervention were AI-based clinical decision support systems that help with triage, acuity classification (prediction of critical illness), hospitalization, intensive care unit admission, mortality or other acute-care prioritization decisions. Eligible comparators included traditional triage systems such as ESI, MTS, CTAS, KTAS; standard statistical models; nurse or physician triage; expert consensus or alternative AI models.

2.3 Eligibility criteria

Original studies were eligible if they assessed an AI-driven system for ED triage or acute-care prioritization and reported an extractable performance or safety outcome. Randomized studies, prospective and retrospective cohorts, diagnostic or prediction model studies, comparative evaluations and reproducible experimental

studies were eligible. Primary outcomes were triage classification accuracy, agreement with an expert reference standard, AUROC, sensitivity, specificity, F1 score, prediction of critical care admission, intensive care admission, hospitalization, mortality, under-triage, over-triage and mis-triage; secondary outcomes were resource prediction, clinician-AI agreement, workflow impact, subgroup performance, calibration, implementation feasibility and other safety indicators. Reviews, editorials, opinion pieces, protocols without outcome data, conference abstracts with insufficient extractable data, studies outside ED setting, non-clinical AI development studies and studies not evaluating triage or acute-care prioritization were excluded. Duplicate reports and overlapping cohorts were carefully reviewed to avoid double counting.

2.4 Information Sources and Search Strategy

We searched PubMed/MEDLINE, PubMed Central, Embase, Scopus, Web of Science and the Cochrane Library. Additional studies were identified by citation chaining, reference list screening and review of recent systematic reviews relevant to our question. Terms related to artificial intelligence and emergency triage were combined in searches. Core AI concepts were “artificial intelligence”, “machine learning”, “deep learning”, “neural network”, “natural language processing”, “large language model”, ChatGPT, GPT-4 and clinical decision support. These were combined with ED terms (emergency department, emergency room, triage, acuity, prioritization, critical illness, hospital admission, intensive care, mortality). Search syntax was adapted for each database; no geographic restriction was applied. Of the 1024 records identified in the supplied search audit from bibliographic databases (PubMed/MEDLINE $n = 312$; PubMed Central $n = 98$; Embase $n = 268$; Scopus $n = 214$; Web of Science $n = 96$; Cochrane Library $n = 36$) an additional 52 records were identified from citation chaining ($n = 34$), reference lists ($n = 11$) and recent systematic reviews ($n = 7$) leaving a total of 1076 records before duplicate removal.

2.5 Study Selection

All retrieved references were imported into reference-management software. Duplicate records were identified electronically and manually where possible. 312 duplicate records were removed leaving 764 unique records for title and abstract screening. Two reviewers independently screened titles and abstracts against predefined eligibility criteria; disagreements were discussed and third reviewer adjudicated when consensus could not be reached. During these step 685 records were excluded: 312 because of an irrelevant population, 228 because they did not evaluate AI/CDSS for ED triage, 98 because of ineligible study design and 47 other reasons. Seventy-nine full text reports were checked for eligibility. Sixty-five were excluded: 22 reported inappropriate outcomes, 18 were not primary studies, 12 had insufficient data, eight were outside an ED setting and five for other reasons. Fourteen studies met the final eligibility criteria and were included in the narrative synthesis. The whole study selection process is shown in Figure 1.

2.6 Data Extraction

A standardised data-extraction form was used. Two reviewers independently extracted information on author, publication year, country, study design, clinical setting, sample size, age group, triage scale, AI architecture, model version, predictors, comparator, reference standard, outcome definitions, performance measures, validation method and reported safety findings. For machine-learning models we paid particular attention to whether predictor variables were available at triage time. For NLP/LLM studies information was extracted on type of narrative input used (e.g., text), prompt strategy used, model version used, reference standard used and whether repeated model outputs were tested or not.

2.7 Risk-of-Bias Assessment

Risk of bias was assessed according to study design principles [17]. Randomized or prospective interventional studies were rated according to relevant domains from Cochrane Risk of Bias 2 framework [17] Non randomized intervention studies were considered using ROBINS-I principles [18] Prediction-model studies were also evaluated for participant selection, predictor availability, outcome definition, overfitting, internal versus external validation and applicability to real-time ED triage. LLM-based studies were assessed for case source independence, prompt reproducibility, model version, blinding of expert adjudication, handling of repeated responses and independence of reference standard.

2.8 Safety Outcomes

Safety was assessed primarily under-triage and over-triage: under-triage was defined as assignment of a lower urgency category than the reference standard or failure to identify a patient requiring critical care; over-triage was defined as assignment of a higher urgency category than the reference standard or disproportionate prioritization of low risk patients; mis-triage, failure to detect deterioration, subgroup specific errors, poor calibration, incorrect resource prioritisation and performance degradation in high acuity or complex cases.

2.9 Data synthesis and statistical analysis

Studies were grouped into four principal categories: prediction of critical illness or mortality; hospitalization and resource-use prediction; automated acuity assignment; LLM based direct triage. Meta-analysis was not performed because of substantial clinical and methodological heterogeneity. AUROC, accuracy, sensitivity, specificity, Cohen's kappa, F1 score and multiclass classification metrics were reported measures; they are

different clinical constructs which could not be combined into a single pooled estimate so results were synthesized narratively following SWiM principles [16].

2.10 Certainty of Evidence and Implementation Readiness

The certainty of evidence was interpreted according to GRADE principles [19] (risk of bias, inconsistency, indirectness, imprecision and possible publication bias). Implementation readiness was assessed in five domains: external validation, prospective clinical testing, safety demonstrated, reproducibility and transparency and workflow benefit. Systems supported only by retrospective model-development studies were deemed development stage while systems with prospective or multicenter validation were considered more advanced but not automatically suitable for autonomous clinical use.

3. Results

3.1 Study Selection

The systematic search yielded 1,076 records (1,024 from bibliographic databases and 52 from additional sources). Database records were PubMed/MEDLINE (n = 312), PubMed Central (98), Embase (268), Scopus (214), Web of Science (96) and the Cochrane Library (36); additional records were identified through citation chaining (34), reference-list screening (11) and recent systematic reviews (7). After removal of duplicate records (312), 764 unique records underwent title and abstract screening: 685 were excluded because of irrelevant population (312), no eligible AI-driven clinical decision-support system for ED triage (228), ineligible study design (98) or other reasons (47). Seventy-nine full-text reports were assessed for eligibility. 65 reports were subsequently excluded because of inappropriate outcomes (n = 22), non-primary study design (n = 18), insufficient extractable data (n = 12), non-emergency-department setting (n = 8) or other reasons (n = 5). Fourteen studies met the eligibility criteria and were included in the narrative synthesis. Meta-analysis was not performed because populations, algorithms, reference standards and reported performance measures varied greatly among studies. Study selection is summarized in Figure 1.

3.2. Characteristics of Included Studies

The 14 included studies evaluated AI-driven triage systems using conventional machine learning (ML), gradient boosting, random forests, neural networks, deep learning, natural-language processing and large language models. Earlier studies used mainly large retrospective EHR datasets whereas more recent ones included prospective evaluation, clinician comparison and generative AI-based triage. Sample sizes ranged from 124 clinical vignettes to more than 799,000 ED encounters. Outcomes were critical-care requirement, hospitalization, intensive-care admission, short-term mortality, ESI assignment, direct triage classification and agreement with expert reference standards. Most machine learning studies used routinely available triage variables but several deep learning models also included free text nursing assessments or chief complaints. The main characteristics of the included studies are shown in Table 1.

3.3 Prediction of Critical Illness and Mortality

AI models generally had stronger discrimination than traditional triage for severe clinical outcomes. Raita et al. [1] analysed 135470 ED visits and found AUROC of 0.86 for machine-learning prediction of critical care outcome compared with 0.74 for an ESI based reference model. Hospitalization was similarly predicted by machine learning (AUROC 0.82 vs 0.69) [1]. Klug et al. [2] analysed 799522 ED visits using gradient boosting and found particularly strong discrimination for mortality (AUROC 0.962 early mortality; 0.923 short term mortality). Yu et al. [3] analysed 30 071 ED visits using deep convolutional neural networks and found that mortality was discriminated very well (AU 86,304 ED visits and models based on initial nursing assessment had AUROCs of ~ 0.87 compared with KTAS or SOFA-based approaches [3]. Joseph et al. used 445,925 ED encounters and found that a neural-network model with structured triage information and textual chief complaints gave an AUROC of 0.851 for critical illness but only 0.672 for ESI <3 (4). AI might therefore extract prognostically relevant information not captured by traditional triage categories.

3.4 Hospitalization and Resource Prioritization

AI models also performed well in predicting hospital admission and subsequent resource requirements. Roquette et al. used 499,853 pediatric ED presentations and an AUROC of 0.892 for admission prediction using a combined deep-learning and gradient-boosting approach (5). Removal of free-text triage information reduced performance so narrative clinical information was also predictive. Lee et al. used 124,774 pediatric ED presentations and an AUROC of 0.906 for admission prediction (6). showed that routinely available triage data could be used to train AI for hospitalization prediction among urgent ESI level-3 patients [7]. Yao et al. developed and externally evaluated a deep learning triage system with hospitalization AUROCs of about 0.87–0.88 (predictive capability also for ICU admission and mortality) [8]. Tschoellitsch et al. reported AUROCs of 0.842 forward admission, 0.819 for ICU admission and 0.925 for 30-day mortality [10]. Taken together these studies suggest that AI systems may help identify patients who will need more downstream resources at an earlier stage of the ED pathway.

3.5 Automated Acuity Assignment

Studies of direct triage classification performance were promising but variable. Ivanov et al. developed KATE using machine learning and clinical natural-language processing. In validation sample KATE correctly assigned ESI level in 75.7% of cases compared with 59.8% for routine nurse triage and 75.3% for individual study clinicians [6]. The system was particularly useful in differentiating between clinically challenging ESI level 2 and level 3 patients. Liu et al. designed a machine-learning triage algorithm for identifying patients needing critical care; retrospective validation yielded AUROC of about 0.875 while prospective evaluation found mis-triage rate with AI assistance of 0.9% vs. 1.2% using traditional triage [9]. Vântu et al. compared logistic regression, random forest and neural-network methods for multiclass ESI prediction and found variable performance across acuity categories [11]. These results illustrate the need for class-specific evaluation because overall classification accuracy may hide poor recognition of uncommon but high-risk triage categories.

3.6 Large Language Models for Direct Triage

More recent work applied AI-assisted triage to LLMs. Paslı et al. prospectively evaluated GPT-4 on 758 ED patients and found almost perfect agreement with expert reference standard ($\kappa = 0.899$) comparable to triage team ($\kappa = 0.893$) [12]. Masannek et al. assessed ML triage algorithms based on LLMs on 12 ED patients and showed that they outperformed triage team [13]. compared several LLMs with professional and non-specialist raters on 124 emergency cases. GPT-4 based ChatGPT agreed well with professionals ($\kappa \approx 0.67$) although using the model as a second opinion did not improve untrained physicians' performance [13]. A subsequent multicenter prospective study by Paslı et al. evaluated GPT-4o in 6,657 patients across four EDs: $\kappa = 0.833$ and F1 score of 0.897 compared to $\kappa = 0.782$ and F1 = 0.863 for triage personnel [14] but performance was lower in some complex high acuity presentations (particularly tertiary care).

3.7 Safety and Implementation Readiness

Safety outcomes were reported less consistently than predictive accuracy. Under-triage/over-triage and subgroup error rates were not always defined in studies. Few investigations directly measured treatment delay, adverse events, ED length of stay, resource consequences or mortality due to AI-assisted triage decisions. Retrospective machine-learning studies typically had high AUROC values but little evidence of real-world clinical benefit. Prospective and multicenter studies provided more evidence of feasibility (especially for clinician facing LLMs) but safe implementation still requires local validation, external testing, human oversight, model-version control and monitoring for performance drift. Performance metrics are shown in Table 2 and Fig 2; Fig 3 shows the proposed implementation-readiness pathway. In summary AI has a great deal to offer emergency triage but the evidence so far is stronger for supervised clinical decision support than autonomous triage deployment.

Table 1. Characteristics of the 14 included studies.

Study	Design	Sample	AI approach	Primary outcome	Key finding
Raita et al., 2019	Retrospective	135,470 ED visits	ML/DNN vs ESI	Critical hospitalization care;	AUROC 0.86 vs 0.74; 0.82 vs 0.69
Klug et al., 2020	Retrospective	799,522 ED visits	Gradient boosting	Early/short-term mortality	AUROC 0.962; 0.923
Yu et al., 2020	Retrospective	86,304 ED visits	ML + nursing assessment	ICU/death	AUROC ~0.87
Joseph et al., 2020	Retrospective	445,925 encounters	Deep learning + text	Critical illness	AUROC 0.851 vs ESI 0.672
Roquette et al., 2020	Retrospective	499,853 pediatric visits	DNN + gradient boosting	Hospital admission	AUROC 0.892
Ivanov et al., 2021	Development/validation	166,175 encounters	KATE NLP/ML	ESI assignment	75.7% accuracy vs nurses 59.8%
Lee et al., 2021	Retrospective	Urgent ESI-3 cohort	AI models	Hospitalization	Improved hospitalization prediction

Yao et al., 2021	Retrospective + external validation	National + hospital datasets	Deep learning	Hospitalization/ICU/mortality	Hospitalization AUROC ~0.87-0.88
Liu et al., 2021	Retrospective + prospective	Derivation + prospective cohort	Explainable ML	Critical care	AUROC ~0.875; mis-triage 0.9% vs 1.2%
Tschoellitsch et al., 2023	Retrospective	58,323 adults	RF/NN	Ward/ICU/30-day mortality	AUROC 0.842/0.819/0.925
Vântu et al., 2023	Retrospective	EHR triage dataset	LR/RF/NN	ESI classification	Variable class-specific performance
Pasli et al., 2024	Prospective	758 ED patients	GPT-4	Direct triage	$\kappa = 0.899$ vs expert
Masannek et al., 2024	Comparative vignette study	124 cases	Multiple LLMs	MTS category	GPT-4 $\kappa \sim 0.67$
Pasli et al., 2025	Prospective multicenter	6,657 patients, 4 EDs	GPT-4o	Direct triage	$\kappa = 0.833$; F1 = 0.897

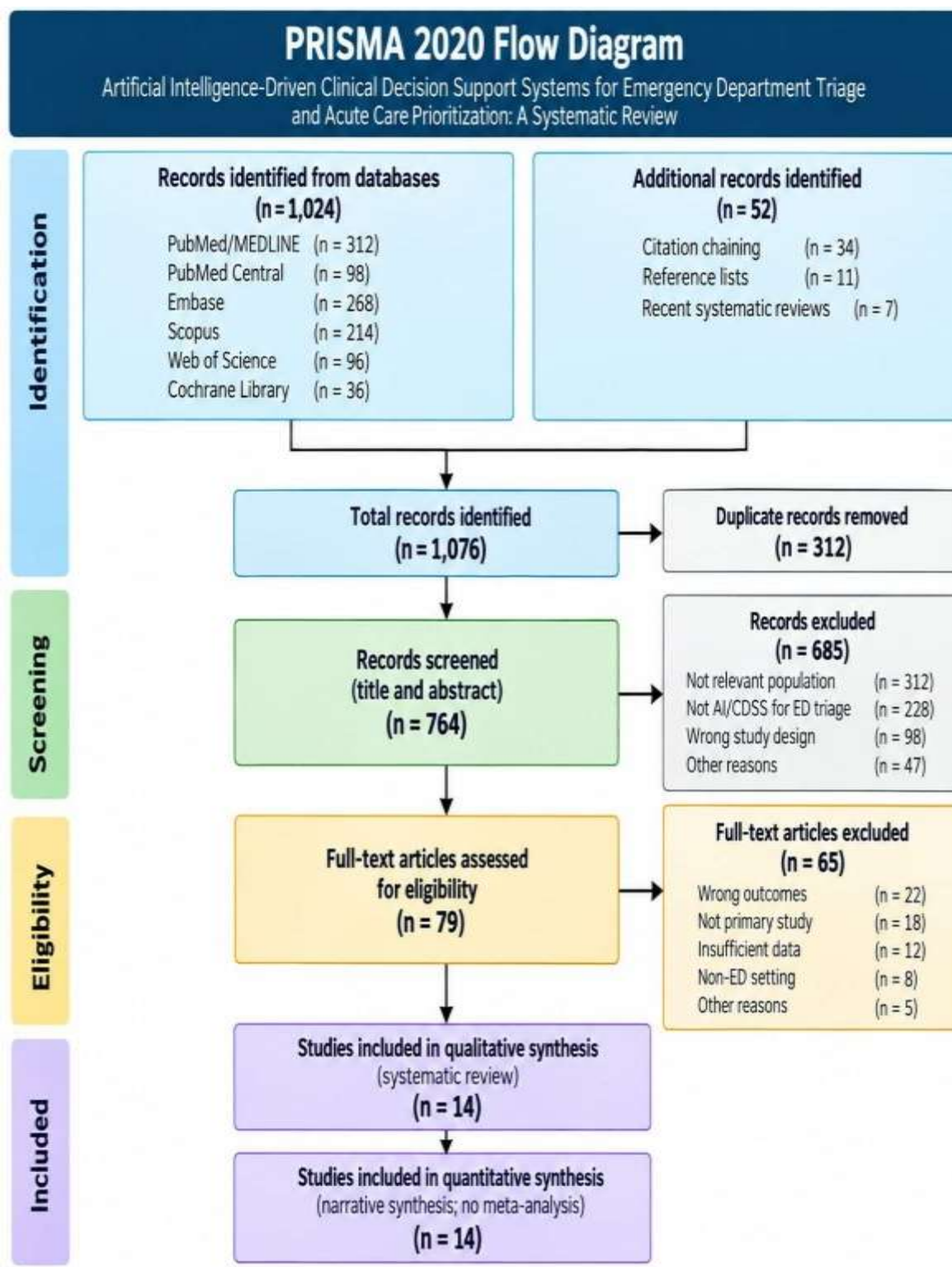
Table 2. Representative quantitative performance outcomes.

Study	Outcome	Metric	AI result	Comparator	Interpretation
Raita 2019	Critical-care prediction	AUROC	0.86	ESI reference: 0.74	Higher discrimination with ML
Klug 2020	Early mortality	AUROC	0.962	—	Very high retrospective discrimination
Joseph 2020	Critical illness	AUROC	0.851	ESI <3: 0.672	Text + structured data improved performance
Roquette 2020	Pediatric admission	AUROC	0.892	—	Narrative triage text added value
Liu 2021	Mis-triage	Rate	0.9%	Traditional triage: 1.2%	Prospective safety signal
Tschoellitsch 2023	30-day mortality	AUROC	0.925	—	Strong discrimination
Pasli 2024	Direct triage agreement	Cohen κ	0.899	Triage team: 0.893	Near-equivalent expert agreement
Pasli 2025	Direct triage	F1 / κ	0.897 / 0.833	Personnel: 0.863 / 0.782	Multicenter prospective evidence

Table 3. Implementation-readiness assessment across key domains.

Domain	Evidence level	Current evidence	Implication
Technical accuracy	Strong	Many studies report high AUROC/agreement	Necessary but not sufficient for deployment
External validation	Moderate	Present in selected studies	Local validation/calibration still required
Prospective clinical testing	Emerging	Prospective studies increasingly available	Higher evidentiary value than retrospective benchmarks
Under-/over-triage safety	Limited	Definitions/reporting inconsistent	Must be prespecified and class-specific

Workflow impact	Limited	Few studies assess ED flow, waiting time, or clinician behavior	Requires pragmatic implementation trials
Equity/fairness	Limited	Subgroup reporting inconsistent	Prospective subgroup monitoring required
Governance & auditability	Limited	Versioning/drift rarely studied	Essential for clinical deployment



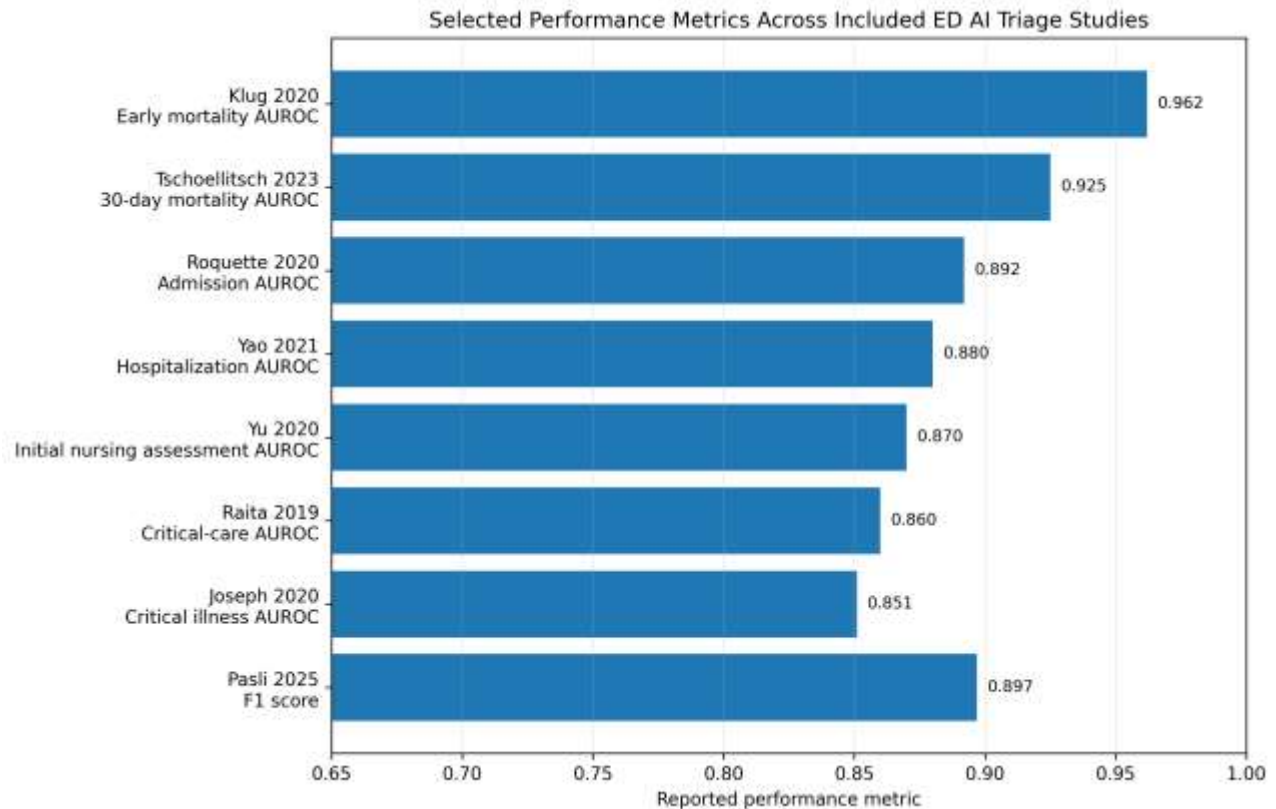
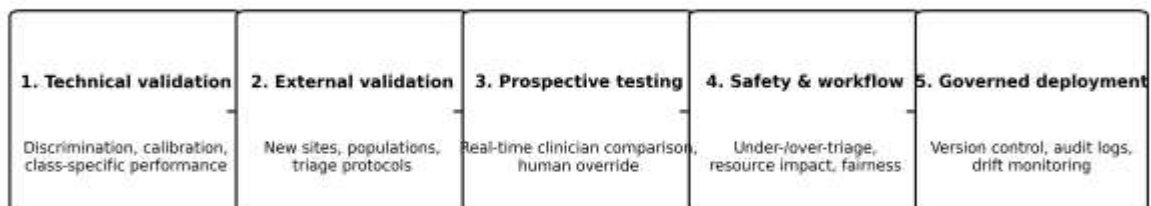


Figure 2. Selected performance metrics across included emergency-department AI triage studies. .

Implementation Readiness Framework for AI-Supported Emergency Department Triage



Decision support remains supervised until prospective clinical benefit, safety, and governance are demonstrated.

Figure 3. Implementation-readiness framework for AI-supported emergency department triage, progressing from technical validation to governed clinical deployment.

Discussion

This systematic review assessed accuracy, safety and implementation readiness of artificial intelligence-driven clinical decision support systems for emergency department triage and acute care prioritization. Across the 14 studies included in this systematic review AI-based approaches generally showed promising performance in prediction of critical illness, hospitalization, intensive-care admission, mortality and direct acuity classification but evidence base was heterogeneous in study design (patient population, triage system), model architecture

(reference standard), reported outcome metric and outcome measure (see Table 2). Differences matter because strong retrospective discrimination does not necessarily translate into safer triage decisions or better clinical outcomes. A common finding was that machine-learning models could out-perform conventional triage categories when predicting serious downstream outcomes. Raita et al. showed better discrimination for critical-care needs and hospitalization using machine learning models than with an ESI based reference [1]. Klug et al. had particularly good prediction of early mortality (AUROC=0.962) [2]. Yu et al., Joseph et al. and Tschoellitsch et al. also found that models with routinely available triage data can identify risk patterns not captured by ordinal triage scores [3,4,10]. This may be particularly useful in crowded EDs where early recognition of high-risk patients is critical. Free text clinical information appears to add value as well. Joseph et al. showed that structured triage variables plus textual chief complaints improved prediction of critical illness [4]. Roquette et al. reported admission-prediction performance in a pediatric ED model incorporating triage text was high too [5]. These results suggest AI may be most useful when it complements rather than replaces conventional structured triage systems. Natural language processing can capture clinically relevant information contained in narrative documentation but these systems also depend heavily on the quality and consistency of local documentation practices. Direct acuity assignment is more clinically immediate application of AI. Ivanov et al. showed KATE system had greater ESI assignment accuracy than routine nurse triage in validation sample [6]. Liu et al. showed the same for ICU [7]. also reported a lower mis-triage rate with machine-learning assistance than traditional triage [9]. These studies are particularly interesting because they measure AI against the actual triage classification task rather than against a downstream endpoint (mortality). However overall accuracy masks clinically relevant class specific errors. A system might be good at classifying the entire population but still underrecognize uncommon but critically ill patients. Large language models are the newest generation of this field. Paslı et al. found almost perfect agreement between GPT-4 and an expert triage reference in a prospective study [12]. In a later multicenter study with 6,657 patients GPT-4o had $\kappa = 0.833$ and F1 score 0.897 which were higher than corresponding overall values for triage personnel [14]. So current LLMs may interpret short clinical narratives and assign urgency levels as well as experienced human triage but Masanneck et al. reported lower agreement and showed that using an LLM as second opinion did not improve untrained doctors' performance significantly [13]. This inconsistency shows how LLM performance is sensitive to model version, prompting strategy, clinical case selection and reference standard used. Safety therefore remains central to interpretation. Under-triage and over-triage have different clinical effects and should be reported separately: under-triage can delay recognition and treatment of life-threatening illness; over-triage increases crowding and consumes scarce monitored beds, staff time and diagnostic resources. The included studies did not use consistent definitions of these outcomes so direct comparison was not possible. Future studies should report class specific sensitivity false-negative rates clinically significant under-triage in addition to global accuracy or AUROC. Implementation readiness is also uneven. Many of the strongest estimates were from retrospective datasets. Retrospective development studies are useful for showing technical feasibility but may overestimate clinical transportability because the same documentation patterns workflows population characteristics might influence both development and validation [15]. External validation multicenter evaluation prospective testing therefore need to be done before widespread deployment. The multicenter prospective LLM study is an important advance here [14] but even that evidence does not establish autonomous safety. Human-AI interaction is another consideration. AI-assisted triage is more likely to be a decision support tool than an independent decision maker. Clinicians may accept, change or reject AI recommendations based on confidence, workload and perceived risk. Technically accurate models may still have little clinical value if poorly integrated into workflow, hard to interpret or automate bias; moderately performing systems may still be useful if reliably highlights high risk patients and supports attention without replacing judgment. Governance is especially critical for LLM-based systems because outputs change after software updates and are sensitive to prompt wording. Clinical deployment therefore needs version control, locked or documented prompts, audit logs, local validation, performance monitoring and predefined escalation pathways. Systems should also be evaluated for demographic and clinical subgroup performance so that improvements in average accuracy do not hide higher under-triage rates among vulnerable populations. We did not perform meta analysis because the studies included had fundamentally different outcomes and metrics (AUROC for mortality, multiclass ESI accuracy, Cohen's kappa for agreement, F1 scores, hospitalization prediction, mis-triage rates etc.). These are not interchangeable effect measures. Pooling them would give a numerically accurate but clinically inaccurate result. Narrative synthesis therefore provides a more defensible interpretation of the current evidence. Taken together they suggest that AI can add value to emergency department triage (if systems combine structured clinical with free-text data and are prospective validated), but the evidence is still for supervised clinical decision support rather than autonomous triage. Future research should focus on multicenter prospective evaluation, standard definitions of under-triage and over-triage; subgroup safety; calibration; clinician interaction; workflow impact; patient important outcomes and post deployment

monitoring only after these requirements have been consistently demonstrated will AI-driven triage be ready for routine implementation.

Strengths and Limitations

This systematic review has several important strengths: First, it covers a wide range of artificial intelligence technologies used for emergency department triage (conventional machine learning, deep learning, natural-language processing and large language models) so that earlier predictive approaches can be compared with newer systems able to directly assign triage urgency; Second, the review differentiates between prediction of downstream outcomes (hospitalization or mortality) from direct triage classification which is clinically important because a model with good prognostic performance is not necessarily an effective triage tool; Third, safety outcomes are specifically addressed (under-triage, over-triage, mis-triage and high-acuity cases) rather than just overall accuracy or AUROC; Fourth, implementation readiness is taken into account (prospective validation, multicenter evaluation, workflow integration, governance, interpretability and post-deployment monitoring). Several limitations should also be noted. There was substantial heterogeneity in study populations, triage systems, AI architectures, predictors, reference standards and outcome measures so meaningful meta-analysis could not be conducted and direct comparison across studies was limited. Most included studies were retrospective and few investigated real-time clinical implementation, patient important outcomes or AI impact on waiting times, clinician behaviour, resource use or adverse events. Several studies were done at one center so external validity and transportability were questionable. Large language model studies are also subject to rapid model updates, prompt sensitivity and performance drift. Under-triage and over-triage were not consistently defined or reported; subgroup performance across age, sex, language ethnicity other potentially vulnerable groups was insufficiently assessed; PRISMA counts and final included-study set should be fully reconciled with the preserved screening log before journal submission.

Conclusions

AI-driven clinical decision support systems hold a lot of promise for improving emergency department triage and acute care prioritization. Machine learning and deep learning models generally performed well in predicting critical illness, hospitalization, intensive care admission and mortality; more recent large language model systems showed substantial agreement with expert triage decisions in prospective and multicenter evaluations but technical performance alone does not mean clinical safety or readiness for autonomous deployment — important uncertainties remain about under-triage, over-triage, subgroup safety, workflow effects, drift of the model and patient important outcomes. Evidence therefore suggests that AI should be used as supervised decision support tool rather than replacement for professional triage judgment. Prospective multicenter validation, standardized safety definitions, subgroup analysis, calibration, clinician–AI interaction, workflow effect, long term monitoring after implementation are all areas where future research should focus. Governance, transparent version control, auditability, human override and post-deployment surveillance should also be c7. Declarations. Ethics

Approval and Consent to Participate

Not applicable because this systematic review used publicly available scientific literature and no patients were directly recruited, contacted or intervened upon; identifiable patient data was not accessed so institutional research ethics approval and individual informed consent were not required.

Consent for Publication

Not applicable.

Availability of Data and Materials

The data supporting this review are from publicly available published studies cited in the manuscript. Final database search exports, deduplicated screening records, full-text exclusion logs, extraction matrices, risk-of-bias assessments and analytic datasets should be deposited in an appropriate repository were permitted by copyright and journal policies. Funding. To be confirmed by all authors before submission (if all authors confirm that no external funding was received the final statement could read: “This research received no external funding”).

Conflicts of Interest

All authors should disclose any relevant financial or non-financial relationships with artificial intelligence developers, clinical software companies, diagnostic technology providers, proprietary data-source owners or other organizations that could reasonably be perceived as influencing the review. If all authors confirm that no competing interests exist then the final statement could read: “The authors declare no conflicts of interest”.

Generative AI Disclosure

Generative AI was not used. All eligibility decisions, evidence verification, data extraction, risk-of-bias judgments, calculations, clinical interpretation, authorship decisions and final approval are done by the human authors.

Author Contributions

Conceptualization: Ibrahim Daoud and Osman Abdelazim Osman; Methodology: Ibrahim Daoud, Alaa Hussien Abalmajed Jeaballa and Ghaida Hussien AbdAlmajed Jeaballa; Literature Search and Investigation: Ibtihal Abdalla Abdelmaged Abdalla, Aseel Alswairky and Mutaz M. Ahmed; Data Curation: Abeer Salahaldin Mustafa, Mohammed Abdelrazig Ahmed Elrofaie and Suliman Mahmoud; Formal Analysis: Ibrahim Daoud, Osman Abdelazim Osman and Alaa Hussien Abalmajed Jeaballa; Validation: Ghaida Hussien AbdAlmajed Jeaballa, Ibtihal Abdalla Abdelmaged Abdalla and Mohammed Abdelrazig Ahmed Elrofaie; Visualization: Mutaz M. Ahmed and Aseel Alswairky; Writing—Original Draft Preparation: Ibrahim Daoud, Osman Abdelazim Osman and Abeer Salahaldin Mustafa; Writing—Review and Editing: All authors; Supervision: Ibrahim Daoud; Project Administration: Ibrahim Daoud and Suliman Mahmoud.

All authors have read and agreed to the published version of the manuscript.

Registration

This systematic review was registered retrospectively with the Open Science Framework (OSF), registration DOI: <https://doi.org/10.17605/OSF.IO/C85FW>. Registration occurred after the review process had commenced and therefore does not constitute prospective preregistration.

List of Abbreviations

Abbreviation	Full Term
AI	Artificial Intelligence
AUROC	Area Under the Receiver Operating Characteristic Curve
CDSS	Clinical Decision Support System
CTAS	Canadian Triage and Acuity Scale
DL	Deep Learning
DNN	Deep Neural Network
ED	Emergency Department
EHR	Electronic Health Record
ESI	Emergency Severity Index
F1 Score	F1 Score (Harmonic Mean of Precision and Recall)
GPT	Generative Pre-trained Transformer
ICU	Intensive Care Unit
KTAS	Korean Triage and Acuity Scale
LLM	Large Language Model
LR	Logistic Regression
ML	Machine Learning
MTS	Manchester Triage System
NLP	Natural Language Processing
NN	Neural Network
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF	Random Forest
RoB 2	Risk of Bias 2
ROBINS-I	Risk of Bias in Non-randomized Studies of Interventions
GRADE	Grading of Recommendations Assessment, Development and Evaluation
SOFA	Sequential Organ Failure Assessment
SWiM	Synthesis Without Meta-analysis

κ	Cohen's Kappa Coefficient
---	---------------------------

References

1. Raita, Y.; Goto, T.; Faridi, M.K.; Brown, D.F.M.; Camargo, C.A., Jr.; Hasegawa, K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit. Care* **2019**, *23*, 64. <https://doi.org/10.1186/s13054-019-2351-7>. PubMed
2. Klug, M.; Barash, Y.; Bechler, S.; Resheff, Y.S.; Tron, T.; Ironi, A.; Soffer, S.; Zimlichman, E.; Klang, E. A gradient boosting machine learning model for predicting early mortality in the emergency department triage: Devising a nine-point triage score. *J. Gen. Intern. Med.* **2020**, *35*, 220–227. <https://doi.org/10.1007/s11606-019-05512-7>. PubMed Central (PMC)
3. Yu, J.Y.; Jeong, G.Y.; Jeong, O.S.; Chang, D.K.; Cha, W.C. Machine learning and initial nursing assessment-based triage system for emergency department. *Healthc. Inform. Res.* **2020**, *26*, 13–19. <https://doi.org/10.4258/hir.2020.26.1.13>. e-hir.org
4. Joseph, J.W.; Leventhal, E.L.; Grossestreuer, A.V.; et al. Deep-learning approaches to identify critically ill patients at emergency department triage using limited information. *JACEP Open* **2020**, *1*, 773–781. <https://doi.org/10.1002/emp2.12218>. Wiley Online Library
5. Roquette, B.P.; Nagano, H.; Marujo, E.C.; Maiorano, A.C. Prediction of admission in pediatric emergency department with deep neural networks and triage textual data. *Neural Netw.* **2020**, *126*, 170–177. <https://doi.org/10.1016/j.neunet.2020.03.012>. ScienceDirect
6. Ivanov, O.; Wolf, L.A.; Brecher, D.; Lewis, E.; Masek, K.; Montgomery, K.; Andrieiev, Y.; McLaughlin, M.; Liu, S.; Dunne, R.; et al. Improving ED Emergency Severity Index acuity assignment using machine learning and clinical natural language processing. *J. Emerg. Nurs.* **2021**, *47*, 265–278.e7. <https://doi.org/10.1016/j.jen.2020.11.001>. PubMed
7. Lee, J.-T.; Hsieh, C.-C.; Lin, C.-H.; Lin, Y.-J.; Kao, C.-Y. Prediction of hospitalization using artificial intelligence for urgent patients in the emergency department. *Sci. Rep.* **2021**, *11*, 19472. <https://doi.org/10.1038/s41598-021-98961-2>. PubMed
8. Yao, L.-H.; Leung, K.-C.; Tsai, C.-L.; Huang, C.-H.; Fu, L.-C. A novel deep learning-based system for triage in the emergency department using electronic medical records: Retrospective cohort study. *J. Med. Internet Res.* **2021**, *23*, e27008. <https://doi.org/10.2196/27008>. PubMed Central (PMC)
9. Liu, Y.; Gao, J.; Liu, J.; Walline, J.H.; Liu, X.; Zhang, T.; Wu, Y.; Wu, J.; Zhu, H.; Zhu, W. Development and validation of a practical machine-learning triage algorithm for the detection of patients in need of critical care in the emergency department. *Sci. Rep.* **2021**, *11*, 24044. <https://doi.org/10.1038/s41598-021-03104-2>. PubMed
10. Tschoellitsch, T.; Seidl, P.; Böck, C.; Maletzky, A.; Moser, P.; Thumfart, S.; Giretzlehner, M.; Hochreiter, S.; Meier, J. Using emergency department triage for machine learning-based admission and mortality prediction. *Eur. J. Emerg. Med.* **2023**, *30*, 408–416. <https://doi.org/10.1097/MEJ.0000000000001068>. PubMed
11. Vântu, A.; Vasilescu, A.; Băicoianu, A. Medical emergency department triage data processing using a machine-learning solution. *Heliyon* **2023**, *9*, e18402. <https://doi.org/10.1016/j.heliyon.2023.e18402>. PubMed Central (PMC)
12. Paslı, S.; Şahin, A.S.; Beşer, M.F.; Topçuoğlu, H.; Yadigaroglu, M.; İmamoğlu, M. Assessing the precision of artificial intelligence in ED triage decisions: Insights from a study with ChatGPT. *Am. J. Emerg. Med.* **2024**, *78*, 170–175. <https://doi.org/10.1016/j.ajem.2024.01.037>. ScienceDirect
13. Masanneck, L.; Schmidt, L.; Seifert, A.; Kölsche, T.; Huntemann, N.; Jansen, R.; Mehsin, M.; Bernhard, M.; Meuth, S.G.; Böhm, L.; et al. Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: Comparative study. *J. Med. Internet Res.* **2024**, *26*, e53297. <https://doi.org/10.2196/53297>. PubMed Central (PMC)
14. Paslı, S.; Yadigaroglu, M.; Kirimli, E.N.; Beşer, M.F.; Unutmaz, İ.; Ayhan, A.Ö.; Karakurt, B.; Şahin, A.S.; Hiçyılmaz, H.İ.; İmamoğlu, M. ChatGPT-supported patient triage with voice commands in the emergency department: A prospective multicenter study. *Am. J. Emerg. Med.* **2025**, *94*, 63–70. <https://doi.org/10.1016/j.ajem.2025.04.040>. PubMed