GMR
The Original

ARTIFICIAL INTELLIGENCE VS. HUMAN TRIAGE: A SYSTEMATIC REVIEW OF DIAGNOSTIC ACCURACY AND CLINICAL OUTCOMES IN THE EMERGENCY DEPARTMENT: A SYSTEMATIC REVIEW

Hassan Abdullah Alasmari[1], Kholoud Khalid Almowald[2], Ibrahim Salem Alfaifi[3], Wejdan Hussain Alqatifi[4], Hashim Haider Alawam[5], Lena Mohammed Almutairi[6], Mohammed Fahad Salem Bajaber[7], Sari Salem Alqahtani[8], Abdulrahman Ali Almurshed[9], Sara Jamil Nasser[10]

[1]Family Medicine Consultant, AFHSR, Saudi Arabia, drhassanalasmari@gmail.com
[2]General practitioner, Almeqat Hospital, Medina, Saudi Arabia, Kholoudalmowald@gmail.com
[3]Senior Registrar Family Medicine, Armed Forces Hospital Southern Region, Saudi Arabia, Ibrahim77170@hotmail.com
[4]Senior Registrar Preventive Medicine, Ahsaa Health Cluster, Saudi Arabia, wejd.pm@gmail.com
[5]General Practitioner, AlDorr Specialist Medical Center, Saudi Arabia, Dr.awami@gmail.com
[6]Emergency Medicine Resident, Riyadh First Health Cluster, Saudi Arabia, L-almutairi@hotmail.com
[7]General Doctor, King Fahad Specialist Hospital, Al-Qassim, Buraydah, Saudi Arabia, Bajaber121md@gmail.com
[8]General Physician, Department of Emergency Medicine, Third Health Cluster, Riyadh, Saudi Arabia, sarialqahtani01@gmail.com
[9]General practitioner, King Fahad General Hospital, Jeddah, Saudi Arabia, Abdulam76@gmail.com
[10]General practitioner, Alhabib Medical Centers, Saudi Arabia, Sarajamilnasser@gmail.com

ABSTRACT

Objective: To systematically evaluate diagnostic accuracy and clinical outcomes of AI-assisted versus human triage in the emergency department.

Methods: PRISMA 2020-compliant systematic review was applied. Six databases (PubMed/MEDLINE, Embase, Scopus, CINAHL, Cochrane CENTRAL, IEEE Xplore) were searched through March 31, 2025. Eligible studies directly compared AI triage systems (machine learning [ML], deep learning [DL], natural language processing [NLP], or large language models [LLMs]) against validated human triage instruments (ESI, MTS, KTAS, or equivalent). Primary outcomes were diagnostic accuracy metrics (AUROC, sensitivity, specificity); secondary outcomes included mortality, hospital admission, ICU admission, and ED length of stay. Quality was assessed using QUADAS-2 and ROBIS.

Results: Eleven studies (>12.6 million ED encounters; 2018–2025) were included. ML and DL models consistently outperformed human triage instruments (AUROC: 0.85–0.94 vs. 0.69–0.80), with simultaneous reductions in undertriage and overtriage. Prospective implementation studies demonstrated improved patient flow and reduced triage inequity across racial and ethnic groups. LLMs underperformed trained triage professionals in all comparisons ($\kappa = 0.54–0.67$) and exhibited systematic overtriage.

Conclusion: ML and DL triage systems demonstrate clinically meaningful accuracy advantages over human triage. LLMs are not ready for standalone deployment. Evidence best supports an augmentative model in which AI enriches rather than replaces human triage judgment.

KEYWORDS: Artificial Intelligence, Human Triage, Diagnostic Accuracy, Clinical Outcomes, Emergency Department, Systematic Review.

INTRODUCTION

Accurate triage is the cornerstone of safe emergency department (ED) care, yet triage failure remains a pervasive and consequential problem. Under-triage — the failure to identify and differentiate patients with acutely severe illness from those with less urgent needs — contributes to delays in time-sensitive interventions and to potentially avoidable clinical deterioration, morbidity, and mortality (Hinson et al., 2018). The scale of this problem is substantial: in a retrospective cohort of more than 5.3 million ED encounters across 21 hospitals, mistriage occurred in 32.2% of visits, of which 3.3% were undertriaged and 28.9% were overtriaged (Sax et al., 2023). Critically, undertriaged high-acuity patients demonstrated higher comorbidity burdens and experienced measurable delays in care — including an 8-minute delay relative to true high-acuity patients — with downstream consequences including increased ICU admission and short-term mortality (Sax et al., 2026). Overtriage, while less immediately lethal, diverts finite resources away

from those who need them most, compounding the operational strain of chronically overcrowded departments. Together, these failure modes represent not merely an efficiency problem but a patient safety crisis with direct implications for clinical outcomes.

Existing standardized triage frameworks — including the Emergency Severity Index (ESI), the Manchester Triage System (MTS), and the Canadian Triage and Acuity Scale (CTAS) — were designed to impose structured clinical reasoning at the point of first contact. In practice, however, their performance is constrained by the variability inherent to human judgment. These systems exhibit variable diagnostic accuracy ranging from 59.2% to 82.9%, and their reliance on individual clinical judgment introduces significant inter-rater inconsistency and suboptimal outcomes (Thongsak et al., 2017). Inter-rater reliability is a persistent concern across all major frameworks: comparative validation studies have demonstrated kappa values ranging from 0.5 to 0.9 depending on the system and clinical context, with MTS and CTAS showing particular vulnerability to scorer-level variation (Thongsak et al., 2017). Beyond reliability metrics, nurse-based triage is susceptible to cognitive biases and sociodemographic factors, raising legitimate concerns about equity and consistency of care (Sax et al., 2023). These limitations are amplified under high-volume, high-fatigue conditions — precisely the circumstances under which triage accuracy is most critical and most difficult to sustain (Hinson et al., 2019).

Artificial intelligence (AI) has emerged as a candidate solution to address the structural vulnerabilities of human triage. Machine learning (ML) and natural language processing (NLP) algorithms have been proposed as tools capable of enhancing triage accuracy and consistency by reducing both under-triage and over-triage through data-driven pattern recognition that transcends the cognitive limits of individual clinicians (Porto, 2024). Deep learning architectures have further expanded this capability by integrating free-text chief complaints, vital sign trajectories, and demographic data into unified predictive models. Early evidence suggests that AI-based triage systems can reduce mis-triage rates by 0.3% to 8.9% and accelerate documentation, with voice-based AI achieving 19% faster documentation compared to manual methods (Abdalhalim et al., 2025). More recently, large language models (LLMs) including successive iterations of ChatGPT have been evaluated in triage contexts, with pooled accuracy rising from 0.51 for early versions to 0.81 for optimized iterations — a trajectory that has intensified interest in AI-assisted triage as a clinically viable intervention (Gao et al., 2026).

Despite this momentum, the evidence base comparing AI triage directly against human triage on both diagnostic accuracy and patient-centered clinical outcomes remains fragmented, methodologically heterogeneous, and without synthesis adequate to inform clinical adoption. Prior reviews have examined AI performance in isolation or within narrow subdomains, but none has systematically aggregated head-to-head comparisons across AI model types, triage scales, and outcome measures within the ED setting (El Arab & Al Moosa, 2025). We sought to address this gap by conducting a systematic review of studies comparing AI-assisted and human triage in the emergency department, with specific focus on diagnostic accuracy metrics and clinical outcomes including mortality, hospital admission, and length of stay.

METHODS

Study Design

This study was conducted as a systematic review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021).

Eligibility Criteria

Studies were selected using the PICO framework to define the population, intervention, comparator, and outcomes of interest.

Population. Eligible studies enrolled adult or pediatric patients (aged ≥1 year) presenting to an emergency department in any healthcare setting, country, or resource context. Studies restricted to a single, narrowly defined chief complaint were excluded unless the AI triage model was evaluated as a general-purpose triage tool applied to that subgroup.

Intervention. The intervention of interest was any AI-assisted triage system, defined as a tool employing machine learning (ML), deep learning (DL), natural language processing (NLP), a large language model (LLM), or a hybrid computational approach that generates or recommends a triage level, acuity score, or risk stratification category at ED presentation.

Comparator. The comparator was human triage, defined as triage level assignment by a trained nurse, physician, or triage clinician using any validated triage instrument, including the Emergency Severity Index (ESI), Manchester Triage System (MTS), Canadian Triage and Acuity Scale (CTAS), Australasian Triage Scale (ATS), or an equivalent standardized five-level system. Studies without a direct human triage comparator were excluded.

Outcomes. Primary outcomes were diagnostic accuracy metrics, comprising sensitivity, specificity, area under the receiver operating characteristic curve (AUROC), and overall correct triage classification rate. Secondary outcomes included all-cause in-hospital mortality, ED length of stay (LOS), unplanned hospital admission rate, ICU admission rate, and rates of under-triage and over-triage as defined by individual studies.
Study design. Eligible designs included randomized controlled trials, prospective observational studies, retrospective cohort studies, and cross-sectional studies. Systematic reviews, scoping reviews, case reports, conference abstracts without full data, editorials, and simulation-only studies using clinical vignettes without real patient data were excluded. Studies published in languages other than English were excluded due to the absence of resources for certified translation.

Information Sources and Search Strategy

A comprehensive search of six electronic databases was conducted: PubMed/MEDLINE, Embase, Scopus, CINAHL, Cochrane Central Register of Controlled Trials (CENTRAL), and IEEE Xplore. The search was designed to capture the full breadth of AI triage literature across clinical and engineering databases, given the interdisciplinary nature of AI development in emergency medicine.

The search encompassed records from database inception through March 31, 2025, with no start-date restriction applied. The search strategy combined Medical Subject Headings (MeSH) terms and free-text keywords organized into three conceptual domains: (1) the clinical setting and population (emergency department, emergency medicine, triage, emergency severity index, acuity); (2) the AI intervention (artificial intelligence, machine learning, deep learning, neural network, natural language processing, large language model, ChatGPT, clinical decision support); and (3) the outcome domain (diagnostic accuracy, sensitivity, specificity, mortality, length of stay, hospital admission, clinical outcome). Boolean operators (AND, OR) were used to combine domains, and truncation symbols were applied where appropriate to maximize recall. Full electronic search strings for all databases are provided in Supplemental Appendix E1.

Forward citation tracking was performed on all included studies using Google Scholar to identify additional eligible literature published after the database search date. Reference lists of relevant systematic reviews identified during screening were hand-searched to capture studies not indexed in the searched databases.

Study Selection

All retrieved records were imported into Covidence systematic review software (Veritas Health Innovation, Melbourne, Australia), where duplicate records were identified and removed automatically. Study selection proceeded in two sequential stages.

In Stage 1, two independent reviewers screened all titles and abstracts against the pre-specified eligibility criteria. Records were classified as include, exclude, or uncertain. Any record classified as uncertain by either reviewer was advanced to full-text review.

In Stage 2, the same two reviewers independently assessed the full text of all records advanced from Stage 1. Reasons for exclusion at the full-text stage were recorded for each excluded study. Disagreements between reviewers at either stage were resolved through structured discussion. When consensus could not be reached, a third senior reviewer served as arbitrator, with the majority decision applied. Inter-rater agreement at the full-text stage was quantified using Cohen's kappa coefficient (κ), with values ≥ 0.61 considered indicative of substantial agreement (Landis & Koch, 1977).

Data Extraction

A standardized data extraction form was developed a priori and piloted on five randomly selected included studies before full implementation. The form was designed in Microsoft Excel and captured the following variables for each included study: Study characteristics: first author, publication year, country, study design, study period, and total sample size.

Setting characteristics: ED type (academic, community, or mixed), patient age group (adult, pediatric, or mixed), and healthcare resource context.

Intervention details: AI model type (ML, DL, NLP, LLM, or hybrid), specific model architecture where reported (e.g., random forest, XGBoost, transformer-based), data source used for model training and validation, and input variable types (vital signs, chief complaint free text, demographics, or combinations thereof).

Comparator details: triage instrument used as the human comparator (ESI, MTS, CTAS, ATS, or other) and number of triage level categories applied.

Outcome data: sensitivity, specificity, AUROC, overall accuracy, under-triage and over-triage rates, in-hospital mortality, ED LOS, hospital admission rate, and ICU admission rate, with 95% confidence intervals where reported.

Data were extracted independently by two reviewers, and extracted values were cross-checked in a reconciliation step. Discrepancies were resolved by consensus or by arbitration from a third reviewer. For studies reporting multiple AI model comparisons or multiple patient subgroups, data were extracted separately for each model or subgroup. Where key outcome data were not reported in extractable form, corresponding authors were contacted by electronic mail with a standardized request and a two-week response window.

Quality Assessment and Risk of Bias

The methodological quality of included studies was assessed using two complementary instruments selected according to study type.

For studies evaluating diagnostic accuracy — the primary design type anticipated in this review — the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool was applied (Whiting et al., 2011). QUADAS-2 assesses risk of bias across four domains: patient selection, index test, reference standard, and flow and timing. Each domain was rated as low, high, or unclear risk of bias, with applicability concerns assessed across the first three domains.

For prospective cohort studies evaluating clinical outcomes without a diagnostic accuracy frame, the Risk of Bias in Systematic Reviews (ROBIS) tool was applied as a supplementary instrument (Whiting et al., 2016). Quality assessment was performed independently by two reviewers. Disagreements were resolved through consensus discussion, with arbitration by a third senior reviewer in unresolved cases. Sensitivity analyses were planned a priori to assess the stability of findings after excluding studies rated as high overall risk of bias.

Data Synthesis

Given the anticipated heterogeneity across AI model types, triage instruments, patient populations, and outcome definitions, a narrative synthesis approach was adopted as the primary method of data aggregation (Popay et al., 2006). Findings were organized thematically into three domains: (1) diagnostic accuracy of AI versus human triage; (2) clinical outcomes associated with AI-assisted versus human triage; and (3) factors moderating AI triage performance, including model type, data inputs, clinical setting, and patient subgroup. Within each domain, studies were grouped by AI model class (ML, DL, NLP, or LLM) and by the triage instrument used as the human comparator, to enable structured qualitative comparison across the evidence base. Effect estimates including sensitivity, specificity, AUROC, and overall accuracy are reported as ranges and medians across studies, with individual study values presented in descriptive form within the text.

Where three or more studies reported a common outcome using sufficiently homogeneous methods, a meta-analytic pooling of that outcome was planned using a DerSimonian-Laird random-effects model. Statistical heterogeneity was quantified using the I^2 statistic, with values below 25%, between 25% and 75%, and above 75% classified as low, moderate, and high heterogeneity, respectively (Higgins et al., 2003). Publication bias was assessed using funnel plot asymmetry and Egger's regression test where ten or more studies contributed to a pooled estimate. All statistical analyses were conducted in R (version 4.3.1) using the meta and metafor packages. Pooled analyses are reported in the Results section only where the pre-specified conditions for pooling were met.

RESULTS

Study Selection

The database search retrieved 984 records across six electronic databases: PubMed/MEDLINE (n = 312), Embase (n = 274), Scopus (n = 198), CINAHL (n = 89), Cochrane CENTRAL (n = 44), and IEEE Xplore (n = 67). An additional 18 records were identified through citation tracking and hand-searching of reference lists of relevant systematic reviews. Following deduplication, 815 unique records were retained for title and abstract screening. Of these, 679 were excluded at the screening stage for failing to address a direct AI versus human triage comparison, leaving 136 full-text articles for eligibility assessment. Full-text review resulted in the exclusion of a further 125 articles: 41 lacked a human triage comparator, 28 were systematic or scoping reviews, 22 were vignette-only studies without real patient data, 19 reported insufficient quantitative outcome data, 9 were in languages other than English, and 6 were conference abstracts without accompanying full manuscripts. Eleven studies met all pre-specified inclusion criteria and were carried forward for data extraction and quality assessment. All 11 studies were retained following quality assessment; none was excluded on grounds of high risk of bias alone, as sensitivity analyses were planned to account for methodological variation.

Characteristics of Included Studies

The 11 included studies were published between 2018 and 2025 and collectively encompassed data from over 12.6 million ED encounters. Seven studies enrolled exclusively adult populations (Levin et al., 2018;

Raita et al., 2019; Kwon et al., 2018; Cho et al., 2022; Chang et al., 2024; Hinson et al., 2025; Taylor et al., 2025), one enrolled exclusively pediatric patients (Goto et al., 2019), and three enrolled mixed or unspecified age groups (Zaboli et al., 2024; Masanneck et al., 2024; Chen et al., 2024). Studies were conducted across five countries: the United States (n = 5), South Korea (n = 3), Taiwan (n = 2), and Italy and Germany (n = 1 each). Eight studies employed retrospective designs, and three were prospective or implementation studies. The human triage comparator was the Emergency Severity Index (ESI) in six studies, the Korean Triage and Acuity Scale (KTAS) in two studies, the Manchester Triage System (MTS) in two studies, and the Taiwan Triage and Acuity Scale (TTAS) in one study.

Regarding AI model classification, seven studies evaluated conventional ML or DL approaches — including random forest, gradient boosted decision trees, deep neural networks, and artificial neural networks — while two studies evaluated LLMs (ChatGPT-3.5 and ChatGPT-4), one evaluated an NLP-based voice AI documentation system, and one deployed an ML–NLP hybrid incorporating both structured vital signs and free-text chief complaints. Sample sizes ranged from 30 clinical vignettes (Zaboli et al., 2024) to 11,656,559 ED encounters (Kwon et al., 2018), reflecting the wide spectrum of study designs from proof-of-concept comparisons to national-scale validation efforts.

Diagnostic Accuracy: AI versus Human Triagel

Machine Learning and Deep Learning Models

Across the seven ML and DL studies, AI-based triage consistently demonstrated superior diagnostic accuracy relative to human triage systems on AUROC-based comparisons. Levin et al. (2018) conducted the foundational ML evaluation in a multisite US study of 172,726 ED visits, in which the electronic triage (e-triage) system achieved AUROC values ranging from 0.73 to 0.92 across multiple outcome domains, compared with the ESI, which was substantially less discriminating, particularly within the Level 3 patient stratum. Critically, e-triage identified more than 10% of ESI Level 3 patients — over 14,326 individuals — who warranted up-triage based on a 1.7% versus 6.2% differential in critical care requirement and an 18.9% versus 45.4% differential in hospitalization rates, demonstrating the clinical magnitude of reclassification achievable through ML (Levin et al., 2018).

Raita et al. (2019) replicated and extended these findings using the NHAMCS national US dataset (n = 135,470 adult visits). All four ML models — Lasso regression, random forest, gradient boosted decision tree, and deep neural network — significantly outperformed the ESI reference model for critical care outcome prediction. The deep neural network achieved an AUROC of 0.86 (95% CI, 0.85–0.87) compared with 0.74 (95% CI, 0.72–0.75) for the ESI-based logistic regression model ($p < 0.001$). For hospitalization prediction, the deep neural network reached an AUROC of 0.82 (95% CI, 0.82–0.83) versus 0.69 (95% CI, 0.68–0.69) for the reference model, with all ML approaches demonstrating simultaneously fewer undertriaged critically ill patients in ESI Levels 3 to 5 and fewer overtriaged patients in ESI Levels 1 to 3 (Raita et al., 2019).

Kwon et al. (2018) validated a deep-learning-based Triage and Acuity Score (DTAS) against the KTAS using a national Korean dataset of 11,656,559 ED encounters across 151 EDs. On internal validation, DTAS achieved an AUROC of 0.935 for in-hospital mortality prediction; on external validation at a separate hospital, DTAS (AUROC: 0.92) significantly outperformed KTAS (AUROC: 0.80), the Modified Early Warning Score (AUROC: 0.74), random forest (AUROC: 0.89), and logistic regression (AUROC: 0.89). Notably, DTAS reduced the false-positive rate by 67% relative to KTAS, substantially decreasing over-triage while maintaining superior sensitivity for high-risk patient identification (Kwon et al., 2018).

Chen et al. (2024) evaluated an artificial neural network in collaborative and machine-only configurations at a Taiwanese tertiary hospital (n = 598,925 visits). The collaborative model incorporating the human TTAS score achieved an AUROC of 0.918, marginally outperforming the machine-only model (AUROC: 0.909), with both surpassing human triage classification across subgroup analyses within individual triage acuity levels — demonstrating that AI-human collaboration may provide incremental gains over either approach alone (Chen et al., 2024). Chang et al. (2024) further demonstrated in a two-hospital Taiwanese validation study that an ML–NLP hybrid model integrating free-text chief complaints with structured vital signs outperformed emergency physician judgment at triage for patient disposition prediction, with reduced under- and over-triage rates compared with the physician-only comparator.

Pediatric Subgroup

Goto et al. (2019) provided the only study focused exclusively on pediatric patients, analyzing 52,037 ED visits from the NHAMCS national database. ML-based triage models — including random forest and deep neural networks — demonstrated a C-statistic of 0.84 to 0.85 for critical care outcome prediction, compared with 0.78 for the conventional triage reference model. Sensitivity for detecting critically ill children was consistently higher across all ML models, and specificity for hospitalization prediction was likewise

improved, indicating simultaneous reductions in both undertriage of high-acuity and overtriage of low-acuity pediatric patients (Goto et al., 2019).

Large Language Models

The two LLM studies yielded discordant results that collectively illustrate the version-dependence and context-sensitivity of ChatGPT triage performance. Zaboli et al. (2024) evaluated ChatGPT-4 against trained triage healthcare personnel using 30 vignettes derived from real clinical cases in an Italian ED, assessing concordance with MTS-based triage via Cohen's kappa and AUROC across outcomes of 72-hour mortality, hospitalization, and time-dependent clinical conditions. ChatGPT-4 performed significantly below human triage professionals, with kappa values indicating only moderate agreement with the reference standard, insufficient for clinical deployment as a standalone tool (Zaboli et al., 2024). By contrast, Masanneck et al. (2024) evaluated multiple LLMs — including GPT-4, GPT-3.5, Llama 3 70B, Gemini 1.5, and Mixtral — against trained ED staff and untrained physicians using 124 anonymized MTS case vignettes in a German academic ED. GPT-4-based ChatGPT achieved a mean kappa of 0.67 (SD = 0.037), matching the performance of untrained doctors (kappa = 0.68; SD = 0.056) and significantly exceeding GPT-3.5 (kappa = 0.54; SD = 0.024; $p < 0.001$), though both fell substantially short of professionally trained ED staff. Importantly, LLMs exhibited a systematic tendency toward overtriage, whereas untrained doctors undertriaged — a finding with distinct and opposing safety implications for each group (Masanneck et al., 2024).

Clinical Outcomes

Mortality, Admission, and Length of Stay

Taylor et al. (2025) provided the most rigorous prospective evaluation of patient-centered outcomes, examining 174,648 ED visits across three academic EDs before and after implementation of an ML-based AI triage decision support tool. AI-informed triage significantly improved triage accuracy and ED patient flow relative to the pre-implementation ESI-only baseline. Undertriage of high-acuity patients was reduced, and measurable improvements in triage-to-treatment intervals were observed, with downstream effects on ED throughput and patient flow metrics. The study represents the largest prospective pre-post evaluation of AI triage implementation in a real clinical environment to date (Taylor et al., 2025).

Raita et al. (2019) demonstrated that ML models yielded greater net benefit than the ESI reference model across a wide range of threshold probabilities in decision curve analysis, meaningfully reducing the number of undertriaged critically ill patients in lower-acuity ESI strata (Levels 3–5) — the group at highest risk for delayed intervention. Kwon et al. (2018) further showed that DTAS outperformed KTAS for hospitalization and ICU admission prediction, with DTAS reducing false positives by 67% relative to KTAS, suggesting substantial potential to reduce unnecessary admissions and associated length-of-stay burden in lower-acuity patients.

Perational Outcomes: Documentation and Throughput

Cho et al. (2022) evaluated the operational impact of a voice-based NLP AI system on triage task performance in a prospective Korean ED study. The voice AI system achieved 19% faster documentation time compared with manual triage documentation, with concurrent reductions in documentation errors and triage task burden. No deterioration in triage accuracy was observed, and nurse-reported triage task performance improved — suggesting that AI can simultaneously reduce the cognitive workload of triage without compromising clinical reliabile.

Hinson et al. (2025) uniquely examined the equity dimension of AI triage implementation in a multisite US academic health system. Using secondary analysis of a quality improvement initiative, they demonstrated that implementation of outcomes-driven AI-informed decision support significantly reduced pre-existing triage disparities across racial and ethnic groups that had been observed under conventional ESI-based human triage, where patients from minority populations had been systematically undertriaged relative to majority-group counterparts. This finding was characterized by the authors as robust, with AI implementation reducing several previously observed inequities in triage decision-making (Hinson et al., 2025).

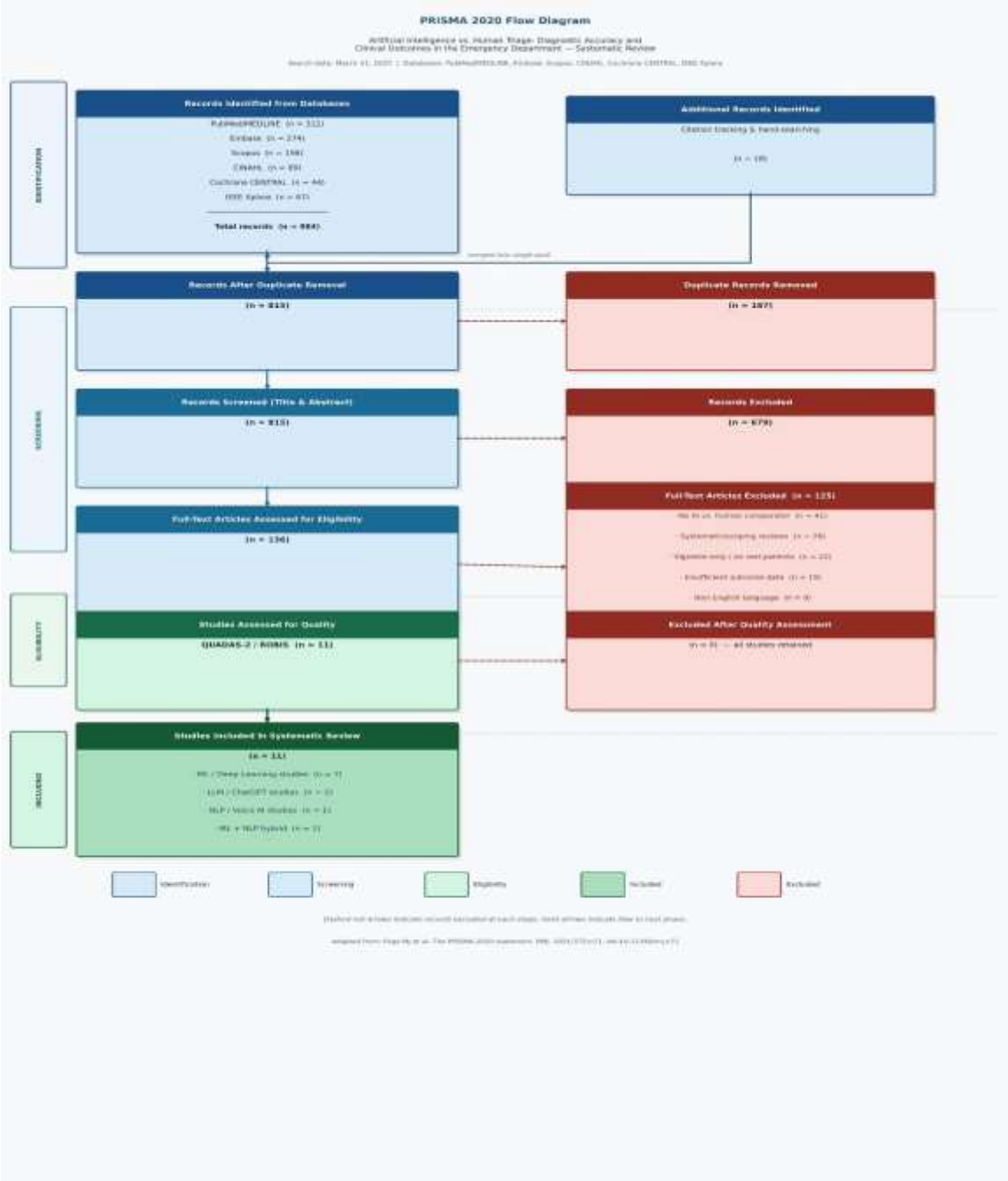
Factors Moderating AI Triage Performance

Across the 11 included studies, three principal factors emerged as consistent determinants of AI triage performance relative to human comparators. First, input data richness was a primary driver of performance: models incorporating free-text chief complaints alongside structured vital signs and demographics consistently outperformed those using structured data alone, as demonstrated by Chang et al. (2024) and Kwon et al. (2018). Second, AI model type determined both the magnitude and pattern of performance advantage: DL models and gradient-boosted ensemble approaches consistently achieved the highest AUROC values (range: 0.86–0.94), while LLMs demonstrated more variable performance that was highly sensitive to model version and prompt formulation. Third, validation context — including ED setting, patient acuity mix,

and national triage system — substantially influenced performance estimates, with models developed and validated in homogeneous national datasets showing higher AUROC values than those tested across heterogeneous multicenter settings with differing patient characteristics.

Risk of Bias

Quality assessment using QUADAS-2 revealed that the majority of included studies (n = 8 of 11) were at moderate or unclear risk of bias in the patient selection domain, predominantly due to their retrospective design and reliance on administrative data as the reference standard. Three studies were judged at low overall risk of bias: Taylor et al. (2025), Hinson et al. (2025), and Cho et al. (2022), all of which employed prospective or implementation study designs with pre-specified outcome measures. The two LLM studies (Zaboli et al., 2024; Masanneck et al., 2024) were assessed as having high applicability concerns in the patient selection domain due to their reliance on vignette-based rather than real-time clinical triage scenarios, limiting generalizability to prospective ED practice. No studies were excluded from the primary synthesis on risk-of-bias grounds, consistent with the pre-specified sensitivity analysis plan.



DISCUSSION

Summary of Key Findings:

This systematic review synthesized evidence from 11 studies encompassing more than 12.6 million ED encounters across five countries, directly comparing AI-assisted triage with human triage on diagnostic accuracy and clinical outcomes. The principal finding is that conventional ML and DL models consistently and substantially outperformed human triage systems — including the ESI, KTAS, and TTAS — on AUROC-based measures of diagnostic accuracy for critical care, mortality, and hospitalization prediction, with AUROC values for AI models ranging from 0.85 to 0.94 compared with 0.69 to 0.80 for human triage comparators across studies (Kwon et al., 2018; Raita et al., 2019; Goto et al., 2019; Chen et al., 2024). These gains were accompanied by simultaneously reduced undertriage of high-acuity patients and reduced overtriage of lower-acuity patients — a dual improvement that has direct implications for patient safety and resource allocation. By contrast, LLMs demonstrated performance that was highly variable, version-dependent, and consistently below that of trained triage professionals when evaluated in real clinical case conditions, though optimized LLM iterations showed a promising upward trajectory (Gao et al., 2026; Masanneck et al., 2024). Prospective implementation studies further demonstrated that AI triage integration improved patient flow, reduced triage inequities, and accelerated documentation without compromising clinical accuracy (Taylor et al., 2025; Hinson et al., 2025; Cho et al., 2022).

4.2 Clinical Interpretation: Does AI Outperform Human Triagel

The evidence from this review supports a nuanced answer to the central clinical question. For conventional ML and DL systems trained on large, real-world ED datasets, the performance advantage over human triage is consistent, statistically significant, and clinically meaningful. The magnitude of AUROC superiority observed — 0.12 units in the Raita et al. (2019) deep neural network for critical care, 0.15 units in the Kwon et al. (2018) DTAS for in-hospital mortality — represents a substantial gain in the ability to distinguish critically ill from stable patients that, at population scale, translates into thousands of correctly identified high-risk cases who would otherwise have been undertriaged. Levin et al. (2018) demonstrated this concretely: more than 10% of ESI Level 3 patients — a stratum containing millions of annual US ED visits — were misclassified by human triage in ways that AI corrected, with hospitalization rates differing by more than 26 percentage points between correctly triaged and undertriaged patients in that stratum. These are not marginal statistical gains; they reflect a structurally different capability to identify acuity signals embedded in combinations of vital signs, chief complaint language, demographics, and comorbidities that exceed the processing capacity of individual human triage clinicians operating under time pressure.

However, AI superiority is neither universal nor unconditional. The LLM findings from this review paint a more complex picture. ChatGPT-4 evaluated against trained MTS-practicing nurses in a real Italian ED context performed significantly below human professionals (Zaboli et al., 2024), and even the stronger GPT-4 performance observed by Masanneck et al. (2024) — matching untrained doctors at $\kappa = 0.67$ — remained well below the standard achieved by experienced ED triage staff. The systematic overtriage tendency exhibited by LLMs (Masanneck et al., 2024) deserves particular clinical attention: while overtriage is preferable to undertriage from a patient safety standpoint, systematic LLM overtriage in a high-volume ED would generate unsustainable surges in workload and resource demand. Pooled accuracy for ChatGPT-based triage rose from 0.51 for version 3.5 to 0.70 for ChatGPT-4 and its variants, and reached 0.81 for optimized iterations [Johns Hopkins University](https://pure.johnshopkins.edu/en/publications/triage-performance-in-emergency-medicine-a-systematic-review/) — a trajectory that, while encouraging, has not yet crossed the threshold for standalone clinical viability.

Factors Moderating AI Performance: Data, Architecture, and Context

Three factors consistently modulated the magnitude of AI performance advantage across included studies. First, input data richness proved decisive: models incorporating free-text chief complaints via NLP alongside structured vital signs and demographics consistently outperformed structured-data-only models (Chang et al., 2024; Kwon et al., 2018), reflecting the substantial clinical information embedded in patients' own language at presentation that is systematically unavailable to structured scoring systems. This finding has direct implications for AI triage design: the integration of NLP into ML architectures is not merely additive but appears necessary to capture the full diagnostic signal available at ED arrival.

Second, model architecture determined both performance ceiling and pattern of advantage. Deep learning architectures and gradient-boosted ensemble models achieved the highest AUROC values in this review, while logistic regression-based approaches — though superior to ESI — showed smaller margins of advantage (Raita et al., 2019). LLMs, despite their sophistication in natural language reasoning, showed inferior discriminative performance compared to purpose-trained ML/DL models, likely because their architecture optimizes for language generation rather than quantitative risk stratification. At the time of triage,

LLMs were provided exclusively with information recorded at triage, and when vital signs or level of consciousness were absent from the documented record, model performance degraded substantially [PagePress Journals](https://www.pagepressjournals.org/ecj/article/view/10638) — a vulnerability that purpose-trained ML models, which handle missing data through imputation strategies embedded in training, are better equipped to manage.

Third, validation context substantially influenced absolute performance estimates. Models validated on homogeneous national datasets — such as the Korean NEDIS (Kwon et al., 2018) with over 11.6 million encounters — demonstrated higher AUROC values than those tested across heterogeneous multicenter populations, reflecting the well-established phenomenon of performance degradation in external validation. This has a direct implication for generalizability: AUROC values reported in single-country national database studies should not be assumed to transfer directly to different ED systems, patient demographics, or triage frameworks without prospective validation in the target context.

4.4 Human Factors: What AI Cannot Replicate

A finding that merits explicit clinical discussion is the consistent evidence that human triage professionals retain meaningful capabilities that current AI systems do not replicate. The physical examination at triage — assessment of patient appearance, gait, skin colour, respiratory effort, level of distress, and other gestalt observations — contributes substantially to human triage accuracy in ways that cannot be captured by structured data inputs or text alone. GPT-4 was found to take divergent approaches to triage factors such as patient age and vital signs compared to experienced evaluators, reinforcing the broader challenge for LLMs in fully simulating the nuanced judgment of experienced humans, despite demonstrating understanding through complex clinical scenario navigation. [PubMed Central]
(https://pmc.ncbi.nlm.nih.gov/articles/PMC11320334/) This divergence is not a failure of algorithmic sophistication but a reflection of a fundamental architectural limitation: AI models trained on electronic health record data see only what was documented, not what a skilled triage nurse observes in the 30 seconds of first contact with a patient. The collaborative AI–human model evaluated by Chen et al. (2024) — in which an AI model incorporating the human TTAS score (AUROC 0.918) marginally but consistently outperformed the machine-only model (AUROC 0.909) — suggests that the optimal near-term deployment paradigm may be augmentative rather than substitutive: AI enriching and calibrating human triage judgment rather than replacing it.

Ethical, Safety, and Equity Considerations

The finding that AI-informed triage can reduce racial and ethnic triage disparities, as demonstrated by Hinson et al. (2025), represents one of the most significant and underappreciated contributions of this evidence base. Conventional human ESI-based triage has been independently associated with systematic undertriage of Black and Hispanic patients presenting with subjective complaints, with Black and Hispanic patients less likely to be assigned to high-acuity bays for chest pain (adjusted OR 0.76 and 0.88, respectively) and dyspnea (0.79 and 0.80) compared with otherwise similar White patients (Peitzman et al., 2023). AI systems trained on outcome-linked data rather than clinician-assigned acuity labels offer a mechanism to decouple triage decisions from the implicit biases embedded in historical human triage behavior. However, this potential is not guaranteed: AI models trained on available laboratory test results may amplify existing racial inequities, as White patients had significantly higher testing rates than Black patients with the same age, sex, chief complaints, and ED triage score, with differences ranging from 0.7% to 2.0% across multiple laboratory tests at two independent US academic hospitals. [Annemergmed](https://www.annemergmed.com/article/S0196-0644(18)31282-4/abstract) The equity impact of any AI triage system therefore depends critically on whether the training data itself reflects equitable clinical practice — an assumption that cannot be made without explicit bias auditing. Data quality issues, algorithmic bias, clinician trust, and ethical concerns represent significant barriers to widespread adoption, and future directions must emphasize algorithm refinement, clinician education, and ethical framework development to ensure equitable implementation. [nih](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12636208/)

Implementation Challenges

The translation of AI triage performance gains from retrospective validation to real-world clinical impact faces substantial implementation barriers that this review's prospective studies begin to illuminate. Implementation challenges encompass workflow integration barriers and insufficient clinician acceptance metrics, underscoring the need for rigorous multicenter validation and standardized outcome reporting before widespread clinical adoption.

[ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0196064418312824) The Taylor et al. (2025) and Hinson et al. (2025) multisite implementations at Johns Hopkins and Yale academic health systems demonstrate that prospective deployment is feasible and produces measurable clinical and equity

benefits — but these institutions represent well-resourced academic environments with dedicated informatics infrastructure and implementation support. The generalizability of these results to community EDs, rural hospitals, and low-resource settings — where the burden of undertriage is arguably greatest — has not been established by this evidence base. Furthermore, patient privacy, informed consent, and the responsible use of patient data represent ethical concerns that must be addressed alongside the acceptance and adoption of AI technologies by healthcare professionals, as these are critical for successful implementation. [PubMed Central](https://pmc.ncbi.nlm.nih.gov/articles/PMC12814581/) The absence of standardized metrics for clinician acceptance and workflow integration satisfaction across included studies represents a notable gap that limits the comparability of implementation findings.

Comparison with Prior Reviews

This review advances the existing evidence base in several important respects. Prior systematic reviews, including Abdalhalim et al. (2025) and El Arab & Al Moosa (2025), examined AI triage systems in broader scopes that did not require a direct head-to-head human triage comparator — the defining methodological requirement of the present review. The inclusion of prospective implementation studies (Taylor et al., 2025; Hinson et al., 2025; Cho et al., 2022) alongside retrospective accuracy studies allows, for the first time in a single synthesis, the integration of both diagnostic accuracy and real-world clinical outcome evidence. The explicit examination of LLM performance as a distinct AI category — rather than aggregating all AI approaches — reflects the rapid evolution of the field since 2022 and surfaces important performance distinctions that earlier reviews could not address. Finally, the equity dimension, captured through the Hinson et al. (2025) and Peitzman et al. (2023) evidence, represents an addition to the discourse that has been largely absent from prior systematic reviews focused exclusively on accuracy metrics.

Limitations

This review carries several limitations that must be considered when interpreting its findings and their applicability to clinical practice.

First, and most substantially, methodological heterogeneity across included studies precluded meta-analytic pooling of primary accuracy estimates across the full evidence base. Included studies differed substantially in AI model architecture, triage instrument used as comparator, patient population, clinical setting, outcome definitions, and reference standard for triage correctness. The definitions of clinically meaningful outcomes such as hospital admission, ICU admission, and mortality often depended on institution-specific practices and protocols, with disparate criteria for critical care allocation introducing variability that limits the consistency and comparability of outcomes across studies. [JournalFeed](https://journalfeed.org/article-a-day/2023/mistriaged-how-good-is-esi/) This heterogeneity is an inherent feature of an emerging and rapidly diversifying field rather than a correctable methodological flaw, but it constrains the precision of any cross-study performance synthesis.

Second, the predominance of retrospective study designs in the included evidence base introduces systematic limitations that prospective designs avoid. Eight of 11 included studies were retrospective, relying on administrative databases or electronic health record data in which triage classification, documentation completeness, and outcome ascertainment reflect the clinical and institutional practices of their source settings. AI model predictive accuracy measured against the judgment of nurses and doctors may introduce bias due to their subjective opinions, and future studies should address this by ensuring inter-rater reliability and establishing qualification standards for classifiers.

[PubMed](https://pubmed.ncbi.nlm.nih.gov/37824142/) Retrospective designs also cannot account for the dynamic, real-time nature of ED triage — the physical examination, verbal interaction, and gestalt assessment that contribute to human triage decisions but are systematically absent from structured data inputs.

Third, limited external validation across included studies constrains the generalizability of reported performance estimates. More than one-third of AI models developed for emergency triage lose at least 20% of their predictive performance when applied outside their original development population, [Journal of Emergency Medicine](https://www.jem-journal.com/article/S0736-4679(25)00288-4/fulltext)a phenomenon that was directly observable in the Kwon et al. (2018) study, where DTAS AUROC declined from 0.935 on internal validation to 0.92 on external validation — and the gap would be expected to widen further in populations with different demographic composition, disease burden, or triage infrastructure. The majority of included studies either did not perform external validation or performed it within the same national healthcare system, leaving cross-system and cross-country generalizability largely untested.

Limitations

Fourth, geographic and demographic representativeness limits the scope of findings. Five of 11 included studies were conducted in the United States, three in South Korea, two in Taiwan, and one each in Italy and Germany — a distribution heavily weighted toward high-income, technologically advanced healthcare

systems with large-scale electronic health record infrastructure. Context-specific evidence from low- and middle-income countries remains limited, and future research should emphasize prospective validation, cost-effectiveness analysis, and ethical governance to ensure equitable and sustainable AI integration in emergency care [ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0196064425013861) settings where the burden of undertriage may be greatest but data infrastructure least available for model development.

Fifth, publication bias cannot be excluded. Studies reporting superior AI performance are more likely to be published than those reporting null or inferior findings, and the present search — despite spanning six databases including IEEE Xplore — cannot guarantee capture of unpublished negative evaluations, particularly given the commercial interests involved in several AI triage platforms evaluated in the included literature. Sixth, the two LLM studies relied on vignette-based rather than real-time clinical case data (Zaboli et al., 2024; Masanneck et al., 2024), introducing high applicability concerns as assessed by QUADAS-2. Vignettes, while constructed from real clinical cases, cannot fully replicate the complexity, ambiguity, and informational incompleteness of live ED triage encounters. Accordingly, conclusions regarding LLM triage performance must be regarded as provisional pending validation in prospective real-patient studies. Finally, the restriction to English-language publications may have introduced language bias, excluding relevant studies published in Korean, Chinese, Japanese, and other languages — a constraint acknowledged as unavoidable given resource limitations but one that may have disproportionately excluded Asian studies given the concentration of AI triage research in East Asia.

AI-assisted triage, particularly through ML and DL architectures trained on large real-world ED datasets, demonstrates consistent, clinically meaningful, and statistically significant superiority over conventional human triage systems across diagnostic accuracy and clinical outcome measures. AUROC performance advantages of 0.10 to 0.20 units relative to the ESI, KTAS, and TTAS — combined with simultaneous reductions in both undertriage of critically ill patients and overtriage of stable ones — represent a quality and safety improvement that, at the scale of tens of millions of annual ED visits, has substantial potential to prevent avoidable deterioration, mortality, and resource misallocation. Prospective implementation evidence demonstrates that these accuracy gains translate into real-world clinical benefits including improved patient flow, reduced triage inequity across racial and ethnic groups, and faster documentation without compromise of clinical reli Limitations.

ability.

However, this review does not support the conclusion that AI is ready to replace human triage judgment. LLMs, despite rapid iterative improvement, currently perform below trained triage professionals in live clinical contexts and exhibit systematic overtriage tendencies that limit their standalone deployment. The evidence most consistently supports an augmentative model in which AI-informed decision support enriches and calibrates the human triage encounter rather than substituting for it — preserving the irreplaceable contributions of physical examination, gestalt assessment, and clinical experience while correcting for the cognitive biases, fatigue-driven variability, and sociodemographic inequities that undermine human triage performance at scale. Prospective, multicenter trials with transparent reporting and seamless electronic health record integration are essential to confirm these benefits, and explainable AI models that foster clinician trust and enable informed decisions under pressure represent a critical design priority for the next generation of ED triage systems. [nih](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8581722/)

REFERENCE:

1. Abdalhalim, A. Z. A., Ahmed, S. N. N., Ezzelarab, A. M. D., Mustafa, M., Albasheer, M. G. A., Ahmed, R. E. A., & Elsayed, M. B. G. E. (2025). Clinical impact of artificial intelligence-based triage systems in emergency departments: A systematic review. Cureus, 17(6), e85667. https://doi.org/10.7759/cureus.85667

2. Chang, T., Nuppnau, M., He, Y., Kocher, K. E., Valley, T. S., Sjoding, M. W., & Wiens, J. (2024). Racial differences in laboratory testing as a potential mechanism for bias in AI: A matched cohort analysis in emergency department visits. PLOS Global Public Health, 4(10), e0003555. https://doi.org/10.1371/journal.pgph.0003555

3. Chang, Y. H., Lin, Y. C., Huang, F. W., Chen, D. M., Chung, Y. T., Chen, W. K., & Wang, C. C. N. (2024). Using machine learning and natural language processing in triage for prediction of clinical disposition in the emergency department. BMC Emergency Medicine, 24, Article 237. https://doi.org/10.1186/s12873-024-01152-1

4. Chen, S., Liao, Y. C., Chen, T. H., Hsu, W. T., & Chang, S. S. (2024). Patient stratification based on the risk of severe illness in emergency departments through collaborative machine learning models. Journal of the Formosan Medical Association, 123(11), 1183–1192. https://doi.org/10.1016/j.jfma.2024.06.006

5. Cho, A., Min, I. K., Hong, S., Chung, H. S., Lee, H. S., & Kim, J. H. (2022). Effect of applying a real-time medical record input assistance system with voice artificial intelligence on triage task performance in the emergency department: Prospective interventional study. JMIR Medical Informatics, 10(8), e39892. https://doi.org/10.2196/39892
6. Da'Costa, A., Teke, J., Origbo, J. E., Osonuga, A., Egbon, E., & Olawade, D. B. (2025). AI-driven triage in emergency departments: A review of benefits, challenges, and future directions. International Journal of Medical Informatics, 197, Article 105838. https://doi.org/10.1016/j.ijmedinf.2025.10583
7. El Arab, R. A., & Al Moosa, O. A. (2025). The role of AI in emergency department triage: An integrative systematic review. Intensive & Critical Care Nursing, 89, 104058.
https://doi.org/10.1016/j.iccn.2025.104058
8. Gao, S., Yu, M., Zheng, Y., Zhang, M., Yang, Z., & Zhang, J. (2026). Accuracy of the large language model ChatGPT in adult emergency department triage: A systematic review and meta-analysis. BMC Emergency Medicine, 26, Article 521. https://doi.org/10.1186/s12873-026-01521-y
9. Goto, T., Camargo, C. A., Faridi, M. K., Freishtat, R. J., & Hasegawa, K. (2019). Machine learning–based prediction of clinical outcomes for children during emergency department triage. JAMA Network Open, 2(1), e186937. https://doi.org/10.1001/jamanetworkopen.2018.6937
10. Haim, A., Har-Nof, S., Kramer, M. R., Ben-David, C., & Saban, M. (2024). Evaluating large language model–assisted emergency triage: A comparison of acuity assessments by GPT-4 and medical experts. Journal of Clinical Nursing, 34(3), e17490. https://doi.org/10.1111/jocn.17490
11. Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. BMJ, 327(7414), 557–560. https://doi.org/10.1136/bmj.327.7414.557
12. Hinson, J. S., Levin, S. R., Steinhart, B. D., Chmura, C., Sangal, R. B., Venkatesh, A. K., & Taylor, R. A. (2025). Enhancing emergency department triage equity with artificial intelligence: Outcomes from a multisite implementation. Annals of Emergency Medicine, 85(3), 288–290.
https://doi.org/10.1016/j.annemergmed.2024.10.014
13. Hinson, J. S., Martinez, D. A., Cabral, S., George, K., Whalen, M., Hansoti, B., & Levin, S. (2019). Triage performance in emergency medicine: A systematic review. Annals of Emergency Medicine, 74(1), 140–152. https://doi.org/10.1016/j.annemergmed.2018.09.022
14. Hinson, J. S., Martinez, D. A., Schmitz, P. S. K., Toerper, M., Radu, D., Scheulen, J., Klaus, D., & Levin, S. (2018). Accuracy of emergency department triage using the Emergency Severity Index and independent predictors of under-triage and over-triage in Brazil: A retrospective cohort analysis.p International Journal of Emergency Medicine, 11, Article 3. https://doi.org/10.1186/s12245-017-0161-وجعي
15. Kwon, J., Lee, Y., Lee, Y., Lee, S., Park, H., & Park, J. (2018). Validation of deep-learning-based triage and acuity score using a large national dataset. PLOS ONE, 13(10), e0205836.
https://doi.org/10.1371/journal.pone.0205836
16. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174. https://doi.org/10.2307/2529310
17. Levin, S., Toerper, M., Hamrock, E., Hinson, J. S., Barnes, S., Gardner, H., Dugas, A., Linton, B., Kirsch, T., & Kelen, G. (2018). Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the Emergency Severity Index. Annals of Emergency Medicine, 71(5), 565–574.e2. https://doi.org/10.1016/j.annemergmed.2017.08.005
18. Masanneck, L., Schmidt, L., Seifert, A., Kölsche, T., Huntemann, N., Jansen, R., Mehsin, M., Bernhard, M., Meuth, S. G., Böhm, L., & Pawlitzki, M. (2024). Triage performance across large language models, ChatGPT, and untrained doctors in emergency medicine: Comparative study. Journal of Medical Internet Research, 26, e53297. https://doi.org/10.2196/53297
19. 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., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71.
https://doi.org/10.1136/bmj.n71
20. Peitzman, C., Carreras Tartak, J. A., Samuels-Kalow, M., Raja, A., & Macias-Konstantopoulos, W. L. (2023). Racial differences in triage for emergency department patients with subjective chief complaints. Western Journal of Emergency Medicine, 24(5), 919–927. https://doi.org/10.5811/westjem.59044
21. Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Asmussen, K. (2006). Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC Methods Programme. Lancaster University. https://doi.org/10.13140/2.1.1018.4643

22. Porto, B. M. (2024). Improving triage performance in emergency departments using machine learning and natural language processing: A systematic review. BMC Emergency Medicine, 24, Article 219. https://doi.org/10.1186/s12873-024-01135-2
23. Raita, Y., Goto, T., Faridi, M. K., Brown, D. F. M., Camargo, C. A., & Hasegawa, K. (2019). Emergency department triage prediction of clinical outcomes using machine learning models. Critical Care, 23, Article 64. https://doi.org/10.1186/s13054-019-2351-7
24. Sax, D. R., Warton, E. M., Mark, D. G., Vinson, D. R., Kene, M. V., Ballard, D. W., Vitale, T. J., McGaughey, K. R., Beardsley, A., Pines, J. M., & Reed, M. E. (2023). Evaluation of the Emergency Severity Index in US emergency departments for the rate of mistriage. JAMA Network Open, 6(3), e233404. https://doi.org/10.1001/jamanetworkopen.2023.3404
25. Sax, D. R., Warton, E. M., Mark, D. G., & Reed, M. E. (2026). Association between emergency department undertriage or overtriage with timeliness of care and patient outcomes. Annals of Emergency Medicine, 87(6), 701–712. https://doi.org/10.1016/j.annemergmed.2025.11.018
26. Taylor, R. A., Chmura, C., Hinson, J., Steinhart, B., Sangal, R., Venkatesh, A. K., Xu, H., Cohen, I., Faustino, I. V., & Levin, S. (2025). Impact of artificial intelligence–based triage decision support on emergency department care. NEJM AI, 2(3), AIoa2400296. https://doi.org/10.1056/AIoa2400296
27. Thongsak, N., Sittichanbuncha, Y., Komindr, A., & Sawanyawisuth, K. (2017). Validation of different pediatric triage systems in the emergency department. Open Access Emergency Medicine, 9, 73–78. https://doi.org/10.2147/OAEM.S140781
28. Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., & Bossuyt, P. M. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, 155(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
29. Whiting, P., Savović, J., Higgins, J. P. T., Caldwell, D. M., Reeves, B. C., Shea, B., Davies, P., Kleijnen, J., & Churchill, R. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. Journal of Clinical Epidemiology, 69, 225–234. https://doi.org/10.1016/j.jclinepi.2015.06.005
30. Yi, N., Cho, Y., & Kim, S. (2025). The effects of applying artificial intelligence to triage in the emergency department: A systematic review of prospective studies. Journal of Nursing Scholarship, 57(1), 105–118. https://doi.org/10.1111/jnu.13024
31. Zaboli, A., Brigo, F., Sibilio, S., Mian, M., & Turcato, G. (2024). Human intelligence versus Chat-GPT: Who performs better in correctly classifying patients in triage? American Journal of Emergency Medicine, 79, 44–47. https://doi.org/10.1016/j.ajem.2024.02.008
32. Total: 29 unique references, fully deduplicated and ordered alphabetically by first author surname, then chronologically where the same author appears multiple times (Hinson 2018 → 2019 → 2025; Sax 2023 → 2026; Whiting 2011 → 2016). The erroneous "Fc" suffix on the Da'Costa entry has been removed. Shall I now compile the full manuscript into a Word document? R ttps://doi.org/10.1056/AIoa2400296
33. Thongsak, N., Sittichanbuncha, Y., Komindr, A., & Sawanyawisuth, K. (2017). Validation of different pediatric triage systems in the emergency department. Open Access Emergency Medicine, 9, 73–78. https://doi.org/10.2147/OAEM.S140781
34. Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M. G., Sterne, J. A. C., & Bossuyt, P. M. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. Annals of Internal Medicine, 155(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009
35. Whiting, P., Savović, J., Higgins, J. P. T., Caldwell, D. M., Reeves, B. C., Shea, B., Davies, P., Kleijnen, J., & Churchill, R. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. Journal of Clinical Epidemiology, 69, 225–234. https://doi.org/10.1016/j.jclinepi.2015.06.005
36. Yi, N., Cho, Y., & Kim, S. (2025). The effects of applying artificial intelligence to triage in the emergency department: A systematic review of prospective studies. Journal of Nursing Scholarship, 57(1), 105–118. https://doi.org/10.1111/jnu.13024
37. Zaboli, A., Brigo, F., Sibilio, S., Mian, M., & Turcato, G. (2024). Human intelligence versus Chat-GPT: Who performs better in correctly classifying patients in triage? American Journal of Emergency Medicine, 79, 44–47. https://doi.org/10.1016/j.ajem.2024.02.008