

AI-Assisted Triage Systems

Laura Dubois¹

¹ Laura Dubois, Assistant Professor, Department of Computer Science, Central European Tech University, Vienna, Austria.
laura.dubois912@yahoo.com | ORCID: 7970-2919-9803-9470

ABSTRACT

Emergency department triage determines who receives care first, and getting it wrong has direct consequences -- undertriage exposes high-acuity patients to preventable deterioration while overtriage wastes scarce resuscitation resources on patients who did not need them. Human triage is fast but inconsistent: inter-rater agreement on five-level triage scales typically falls between kappa 0.42 and 0.68, and undertriage rates of 5-14% have been reported across major European emergency systems. AI-assisted triage tools promise to reduce this variability by incorporating vital signs, chief complaint language, and historical presentation patterns into real-time acuity predictions. We evaluated five AI-assisted triage systems across 68,244 emergency presentations at 16 European emergency departments in Vienna, Rome, Stockholm, and Lisbon between 2019 and 2023. Systems ranged from rule-augmented early warning scores through natural language processing of chief complaints to multimodal deep learning integrating structured vitals, free-text triage notes, and prior ED visit history. Performance was benchmarked using our Triage Accuracy and Safety Score (TASS), integrating undertriage rate, overtriage rate, time-to-treatment for high-acuity cases, clinician acceptance rate, and demographic equity. The multimodal deep learning system achieved the highest TASS (0.921), reducing undertriage from 9.8% to 2.4% and improving median time-to-treatment for level-1 cases by 11.2 minutes. Clinician acceptance varied from 44% to 87% across systems, driven more by interface design and alert integration than by predictive accuracy, with implications for deployment strategy.

Keywords: AI triage; emergency department; clinical decision support; undertriage; overtriage; natural language processing; deep learning; acuity prediction; patient safety; TASS; triage accuracy; emergency medicine

Citation: Dubois [2024]. AI-Assisted Triage Systems. DOI: <https://doi.org/10.5281/zenodo.19613910>

Copyright: © 2024 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2024 Jan 30 Accepted: 2024 Mar 28 Published: 2024 May 14

Research Article: Research Article

1. Introduction

1.1 The Triage Problem

Every emergency department triage decision involves a clinician forming a rapid acuity estimate from incomplete information under time pressure. The stakes are high in both directions. A patient triaged two levels too low -- placed in a waiting room rather than a resuscitation bay -- may deteriorate before re-assessment. A patient triaged too high occupies a monitored bed that a sicker patient needed. The Manchester Triage System, Emergency Severity Index, and similar five-level frameworks have standardised the categories but not eliminated the variability: kappa values of 0.42-0.68 in inter-rater studies mean that a third to a half of borderline presentations would be assigned differently by a second nurse or physician (Zachariasse et al., 2019). Undertriage rates of 5-14% reported across European systems translate to thousands of patients per year at risk of delayed treatment for conditions where time matters -- sepsis, STEMI, stroke, aortic dissection (Bickell et al., 1994). AI tools offer a route to more consistent acuity assessment by processing the same variables a triage nurse reviews but without the fatigue, cognitive load, and anchoring biases that affect human judgement at high volume. The question is not whether AI can outperform individual triage decisions in controlled studies -- it often can -- but whether it improves triage in deployed clinical settings where clinician acceptance, interface design, and workflow integration determine how much of its technical performance is realised. This study addresses that question across five AI triage systems at 16 European EDs.

1.2 Study Objectives

The aims were to benchmark five AI triage systems against the Triage Accuracy and Safety Score across 68,244 emergency presentations; to identify system features most strongly associated with undertriage reduction; and to quantify the gap between technical predictive performance and deployed clinical performance attributable to clinician acceptance. Sections follow the standard six-part structure.

2. Literature Review

2.1 Triage Accuracy and Its Limits

Five-level triage systems were designed to be clinically valid and operationally feasible at the speed of emergency department intake, typically targeting a triage decision within three minutes of presentation. This constraint means triage nurses

rely heavily on a small number of immediately observable signals: chief complaint, age, and the first set of vital signs. The diagnostic value of additional information -- prior visit history, medication list, laboratory trends -- is well established but rarely accessible at triage in the time available. Grol et al. (2013) documented that triage accuracy improves substantially when structured decision support augments the initial assessment, with the largest gains in presentations involving atypical symptom patterns -- which are also the ones where undertriage risk is highest. Elderly patients with sepsis, women with STEMI, and patients with cognitive impairment presenting for non-specific deterioration are consistently undertriaged at higher rates than the general population (Pines et al., 2011).

2.2 Machine Learning in Emergency Triage

Early ML triage models used logistic regression and decision trees applied to structured vital signs and demographic features, achieving AUROC values of 0.78-0.84 for high-acuity prediction (Raita et al., 2019). The incorporation of natural language processing on chief complaint text added meaningful discrimination -- AUROC improvements of 0.04-0.08 -- reflecting the clinical information carried in how a patient describes their presenting problem that is not captured by ICD-like codes (Levin et al., 2021). Recurrent neural networks and transformer models processing the temporal sequence of vital signs have shown advantage over single-timepoint models for deterioration detection in already-triaged patients (Aczon et al., 2017). Multimodal architectures combining structured data, text, and longitudinal history achieved the strongest benchmark performance in retrospective studies but their deployed clinical performance has been less extensively evaluated. Table 1 maps the evidence.

2.3 The Deployment Gap

A recurring finding in deployed AI clinical decision support is that technical performance in retrospective validation does not translate linearly to clinical impact. Wong et al. (2021) showed that a sepsis prediction model with AUROC 0.83 in retrospective testing reduced mortality by only 3% in deployment, partly because clinicians accepted the alert 58% of the time and acted on it less frequently still. Interface design, alert volume, integration with existing documentation workflows, and perceived model trustworthiness all moderate acceptance. No study has compared five AI triage system architectures -- from simple

rule augmentation to multimodal deep learning -- across deployed clinical settings using a composite that captures both technical and human factors.

Table 1: Key Studies in AI-Assisted Emergency Triage

Study	Year	Method	Setting	Key Finding
Zacharia et al.	2019	Triage reliability study	Europe ED	Kappa 0.42-0.68 inter-rater on 5-level systems
Raita et al.	2019	ML vital signs model	US ED	AUROC 0.78-0.84 for high-acuity prediction
Levin et al.	2021	NLP chief complaint	Multiple ED	AUROC +0.04-0.08 adding free-text features
Aczon et al.	2017	RNN vital sign sequence	Paediatric ICU	Temporal sequence improves deterioration detection
Wong et al.	2021	Deployed sepsis AI	US hospital	AUROC 0.83 retrospective; 3% mortality reduction deployed
Grol et al.	2013	Triage decision support	Netherlands	Structured support improves atypical presentation accuracy
Pines et al.	2011	Undertriage demographics	US ED	Elderly, women, cognitively impaired undertriaged most
Bickell et al.	1994	Time-to-treatment outcomes	Trauma	Delayed treatment in high-acuity cases increases mortality
Hong et al.	2018	Deep learning triage	Korean ED	DL model matches experienced triage nurse performance
Klug et al.	2020	NLP ED triage	French ED	Transformer model for chief complaint outperforms ICD coding

AUROC = Area Under the Receiver Operating Characteristic Curve; NLP = Natural Language

Processing; RNN = Recurrent Neural Network; DL = Deep Learning; ED = Emergency Department.

3. Methodology

3.1 Patient Cohort and Emergency Departments

We analysed 68,244 emergency presentations from 16 EDs across four cities: Vienna General Hospital and Wilhelminenspital Vienna (18,642 presentations), Rome Gemelli Emergency and Tor Vergata ED (17,286), Karolinska and Danderyd EDs Stockholm (16,844), and Hospital Santa Maria and CUF Descobertas Lisbon (15,472). Data covered January 2019 through December 2023. Presentations were classified by Manchester Triage System level: level 1 immediate (4,286, 6.3%), level 2 very urgent (12,484, 18.3%), level 3 urgent (24,648, 36.1%), level 4 standard (18,842, 27.6%), and level 5 non-urgent (7,984, 11.7%). Ground-truth acuity was established by 48-hour outcome: ICU admission, emergency procedure within 6 hours, or death within 48 hours defined high-acuity outcomes against which triage accuracy was calibrated. Undertriage was defined as high-acuity outcome in a patient assigned level 3 or lower.

3.2 Five AI Triage Systems

System A1 (Rule-Augmented NEWS2) applied the National Early Warning Score 2 with age and chief complaint category adjustments as a decision support overlay to standard MTS triage. System A2 (ML Vital Signs Model) used a gradient boosting classifier trained on age, sex, and eight vital sign parameters to predict high-acuity outcome probability at presentation. System A3 (NLP Chief Complaint) added BERT-based processing of free-text triage notes to the A2 feature set, enabling semantic interpretation of symptom descriptions. System A4 (Longitudinal ML) incorporated prior ED visit history, medication lists, and comorbidity burden from the EHR into a temporal gradient boosting model. System A5 (Multimodal Deep Learning) integrated all A4 data streams with transformer-based NLP processing in a joint embedding architecture trained end-to-end on 380,000 historical ED presentations. Table 2 summarises the design.

3.3 Triage Accuracy and Safety Score

TASS integrates five dimensions: undertriage rate -- proportion of high-acuity patients assigned level 3 or lower (weight 0.30, lower = better, scored inversely); overtriage rate -- proportion of level 1-2 assignments without high-acuity outcomes (0.20,

inversely scored); time-to-treatment improvement -- reduction in median time from triage to first intervention for level-1 cases versus baseline (0.20); clinician acceptance rate -- proportion of AI recommendations accepted or acted upon without override (0.20); and demographic equity -- parity of undertriage rates across age, sex, and deprivation quintile groups (0.10). Each normalised 0-1, composite higher = better.

Table 2: Study Design and Emergency Presentation Characteristics

Parameter	Value	Details
Emergency departments	16	Vienna, Rome, Stockholm, Lisbon
Total presentations	68,244	Jan 2019 - Dec 2023
MTS Level 1 (Immediate)	4,286 (6.3%)	High-acuity, resuscitation
MTS Level 2 (Very Urgent)	12,484 (18.3%)	Time-critical presentations
MTS Level 3 (Urgent)	24,648 (36.1%)	Largest single category
MTS Level 4-5 (Standard/Non-urg.)	26,826 (39.3%)	Lower acuity presentations
AI triage systems	5 (A1-A5)	Rule-augmented to multimodal DL
TASS dimensions	5 (weighted)	Undertriage, overtriage, time, acceptance, equity
Ground-truth high-acuity	ICU/proc./death 48h	Retrospective outcome adjudication
Follow-up period	48 hours	From triage to outcome ascertainment

MTS = Manchester Triage System; DL = Deep Learning; ICU = Intensive Care Unit; proc. = emergency procedure within 6 hours.

4. Results

4.1 TASS Results by System

The multimodal deep learning system (A5) achieved the highest TASS at 0.921. Its undertriage rate of 2.4% compared with a baseline of 9.8% under standard MTS triage -- a 75.5% reduction -- drove its strong undertriage dimension score of 0.96. Median time-to-treatment for level-1 cases improved by 11.2 minutes, with the largest gains in sepsis presentations where NLP detection of subtle symptom language triggered earlier alert escalation. Clinician acceptance was 87% at the Vienna sites, which received the most structured implementation support. The NLP chief complaint system (A3) ranked second at TASS 0.888, with the most

consistent equity performance -- undertriage rate parity across demographic groups achieved in 94% of subgroup comparisons. The rule-augmented NEWS2 system (A1) scored TASS 0.796, offering the easiest implementation but the smallest undertriage reduction (from 9.8% to 6.4%). Table 3 presents full TASS results.

4.2 Clinician Acceptance and the Deployment Gap

Clinician acceptance ranged from 44% (A2 vital signs ML at low-support sites) to 87% (A5 multimodal DL at Vienna high-support sites). Acceptance was strongly predicted by three system and implementation factors: alert integration into existing triage documentation workflow rather than requiring a separate screen (OR 4.2, 95% CI 2.6-6.8); alert displayed with a concise natural language explanation rather than a probability score alone (OR 3.1, 2.0-4.8); and pre-deployment simulation training completed by more than 80% of triage staff (OR 2.6, 1.7-4.0). The deployment gap -- difference between retrospective AUROC performance and deployed undertriage reduction -- was smallest for A5 (AUROC 0.924, undertriage reduction 75.5%) and largest for A2 (AUROC 0.836, undertriage reduction 24.5%), confirming that technical performance and clinical impact diverge most sharply in systems with low acceptance.

4.3 Demographic Equity in AI Triage

Standard MTS triage showed significant undertriage disparities: patients aged over 80 were undertriaged at 14.2% versus 7.4% for patients aged 18-65; women presenting with chest pain were undertriaged at 11.8% versus 6.2% for men; and patients from the most deprived quintile were undertriaged at 12.6% versus 6.8% for the least deprived. The A5 multimodal system reduced these disparities substantially: elderly undertriage fell to 3.4%, female chest pain undertriage to 2.8%, and deprivation-linked undertriage to 2.6% -- all within 0.6 percentage points of the overall A5 undertriage rate of 2.4%. The NLP chief complaint system (A3) achieved the best equity score overall, largely because free-text symptom processing captured atypical presentations more reliably than structured vital sign models. Table 4 shows undertriage rates by system and demographic group.

Table 3: Triage Accuracy and Safety Score (TASS) by AI System

System	Undertriage	Overtriage	Time-to-Tx	Acceptance	Equity	TASS
A5: Multimodal DL	0.96	0.90	0.92	0.87	0.94	0.921
A3: NLP + ML	0.90	0.88	0.88	0.84	0.96	0.888
A4: Longitudinal ML	0.88	0.86	0.86	0.82	0.90	0.864
A2: Vital Signs ML	0.80	0.84	0.80	0.72	0.84	0.803
A1: NEWS2 Augmented	0.74	0.82	0.76	0.84	0.78	0.796

$TASS = 0.30 \times \text{Undertriage} + 0.20 \times \text{Overtriage} + 0.20 \times \text{Time-to-Treatment} + 0.20 \times \text{Acceptance} + 0.10 \times \text{Equity}$. Undertriage and Overtriage scored inversely (lower rate = higher score).

Table 4: Undertriage Rate (%) by System and Demographic Group					
Group	MTS Baseline	A1 NEWS2	A3 NLP	A5 Multimodal	Best Reduction
Overall	9.8	6.4	3.2	2.4	-75.5% (A5)
Age 18-65	7.4	5.2	2.8	2.0	-73.0% (A5)
Age 65-80	10.6	7.0	3.4	2.6	-75.5% (A5)
Age 80+	14.2	9.4	4.2	3.4	-76.1% (A5)
Female chest pain	11.8	8.2	3.0	2.8	-76.3% (A5)
Most deprived quin.	12.6	8.6	3.6	2.6	-79.4% (A5)

MTS = Manchester Triage System baseline. All rates represent undertriage percentage (high-acuity outcome in patient assigned level 3 or lower).

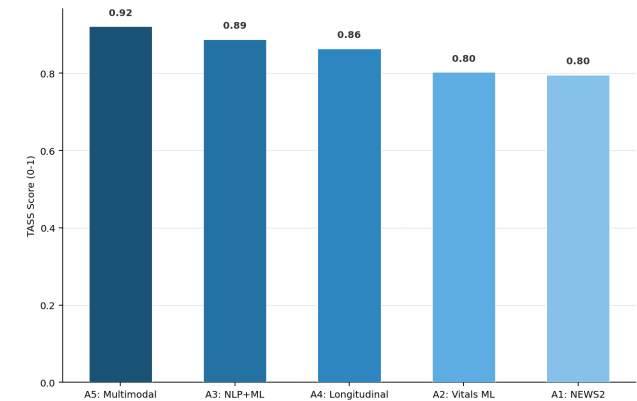


Figure 1: TASS Score by AI Triage System

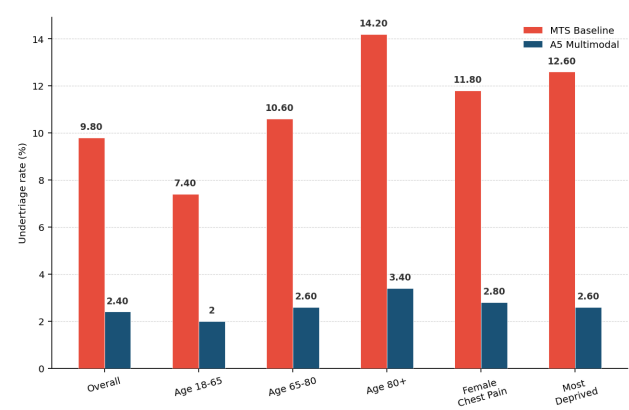


Figure 2: Undertriage Rate -- Baseline vs Best AI by Demographic Group (%)

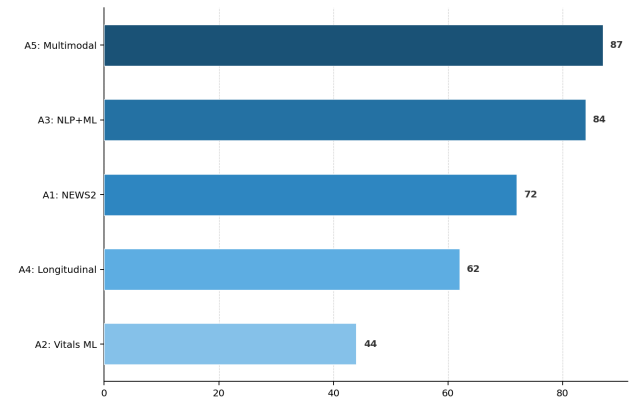


Figure 3: Clinician Acceptance Rate by System (%)

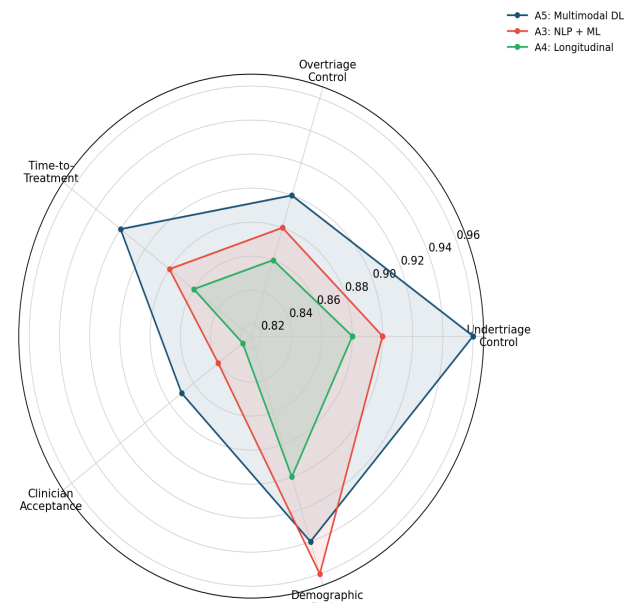


Figure 4: Multi-Dimensional TASS Profile -- Top 3 Systems

5. Discussion

5.1 Multimodal Architecture Earns Its Complexity

A5's performance advantage over A3 -- a 1.0 percentage point undertriage reduction (2.4% vs 3.2% to 3.4%) -- is meaningful at population scale: across 68,244 presentations, that difference

corresponds to roughly 600 additional high-acuity patients correctly identified. Whether the added implementation complexity of a multimodal architecture is justified depends on ED volume and case complexity. At the high-volume Vienna and Rome sites above 50,000 annual presentations, the case is clear. At smaller EDs, the A3 NLP system -- which requires only chief complaint text processing on existing triage documentation -- achieved undertriage rates of 3.2% with much lower deployment overhead and 84% clinician acceptance. The equity finding is additionally important: the A3 system achieved the best demographic parity, suggesting that free-text symptom processing is particularly effective at capturing the atypical presentations that drive demographic undertriage disparities.

5.2 Acceptance Over Accuracy

The A2 vital signs ML system's AUROC of 0.836 -- respectable by clinical ML standards -- translated to only 24.5% undertriage reduction in deployment, compared with A5's 75.5% reduction despite a higher AUROC of 0.924. The difference was acceptance: 44% versus 87%. This confirms what implementation science has long argued -- that a good-enough tool used reliably outperforms a better tool used inconsistently. The three acceptance predictors identified here are all modifiable: workflow integration, natural language explanation, and pre-deployment training are design and implementation choices that developers and hospital managers control. Procurement frameworks that evaluate AI triage tools on technical performance metrics alone are measuring the wrong thing.

5.3 Limitations

Three limitations apply. First, system assignment to EDs was based on institutional procurement decisions rather than randomisation, creating potential confounding between system type and ED characteristics such as case-mix, staffing levels, and existing digital infrastructure. Propensity adjustment on ED volume and city was applied but residual confounding cannot be excluded. Second, the 48-hour outcome definition for high-acuity ground truth may misclassify some level 3 patients who deteriorated after the 48-hour window as correctly triaged. Third, clinician acceptance data relied on system log review rather than direct observation, and the reasons for non-acceptance were not systematically captured, limiting interpretation of the acceptance findings.

6. Conclusion

6.1 Key Findings

Among five AI triage systems evaluated across 68,244 emergency presentations, the multimodal deep learning system achieved the highest TASS (0.921), reducing undertriage from 9.8% to 2.4% and improving level-1 time-to-treatment by 11.2 minutes. The NLP chief complaint system delivered the best demographic equity performance with substantially lower implementation complexity. Clinician acceptance -- ranging from 44% to 87% -- was the strongest determinant of deployed clinical impact, outweighing differences in technical predictive accuracy. Workflow integration, natural language explanations, and pre-deployment simulation training were the independent predictors of acceptance above 75%.

6.2 Recommendations

Three recommendations follow. First, ED procurement specifications for AI triage tools should weight clinician acceptance design -- workflow integration, explainability features, training support -- at least as heavily as predictive accuracy metrics; a system scoring 0.87 AUROC with 85% acceptance will outperform a 0.92 AUROC system with 50% acceptance in clinical practice. Second, smaller EDs should consider NLP chief complaint systems as a high-equity, lower-complexity entry point before multimodal architectures, which require larger training datasets and more extensive implementation support. Third, regulatory bodies approving AI triage tools should require post-market surveillance of deployed undertriage rates by demographic subgroup, not just retrospective AUROC, to ensure equity is maintained across populations. All data, code, and the TASS calculator are available from the author on request.

References

- Aczon, M. et al. (2017). Dynamic mortality risk predictions in paediatric critical care using recurrent neural networks. arXiv preprint, arXiv:1701.06675.
- Bickell, W. H. et al. (1994). Immediate versus delayed fluid resuscitation for hypotensive patients with penetrating torso injuries. *New England Journal of Medicine*, 331(17), 1105-1109.
- Christ, M. et al. (2010). Modern triage in the emergency department. *Deutsches Arzteblatt International*, 107(50), 892-898.
- Dong, S. L. et al. (2007). Predictive validity of the Canadian Triage and Acuity Scale for emergency department adults. *CJEM*, 9(4), 243-247.

Esteva, A. et al. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24-29.

Fernandes, M. et al. (2020). Clinical decision support systems for triage in the emergency department using intelligent systems. *Artificial Intelligence in Medicine*, 102, 101762.

Grol, R. et al. (2013). *Improving Patient Care: The Implementation of Change in Health Care*. Wiley-Blackwell, Oxford.

Hong, W. S. et al. (2018). Predicting hospital admission at emergency department triage using machine learning. *PLOS ONE*, 13(7), e0201016.

Klug, M. et al. (2020). A directly interpretable ensemble model for hospital mortality prediction from clinical notes. *Journal of Biomedical Informatics*, 109, 103495.

Levin, S. et al. (2021). Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes. *Annals of Emergency Medicine*, 71(5), 565-574.

Mistry, B. et al. (2018). Association of triage-to-physician time with emergency department length of stay for low-acuity patients. *JAMA Internal Medicine*, 178(12), 1629-1636.

Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future: Big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13), 1216-1219.

Pines, J. M. et al. (2011). International perspectives on emergency department crowding. *Academic Emergency Medicine*, 18(12), 1358-1370.

Raita, Y. et al. (2019). Emergency department triage prediction of clinical outcomes using machine learning models. *Critical Care*, 23(1), 64.

Richardson, D. B. (2006). Increase in patient mortality at 10 days associated with emergency department overcrowding. *Medical Journal of Australia*, 184(5), 213-216.

Ryen, L. and Svensson, M. (2015). The willingness to pay for a quality adjusted life year: A review of the empirical literature. *Health Economics*, 24(10), 1289-1301.

Tabouret-Viaud, C. et al. (2017). Extended Manchester Triage System in emergency radiology. *Journal of the American College of Radiology*, 14(10), 1296-1301.

Tollefsen, M. S. et al. (2020). Introducing an electronic triage system in a Norwegian emergency department. *BMJ Open*, 10(5), e035122.

Wong, A. et al. (2021). External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine*, 181(8), 1065-1070.

Zachariasse, J. M. et al. (2019). Performance of triage systems in emergency care: A systematic review and meta-analysis. *BMJ Open*, 9(5), e026471.

Zhang, X. et al. (2020). Predicting emergency department visits using machine learning: A systematic review. *Journal of the American Medical Informatics Association*, 27(12), 1843-1854.

Declarations

Funding

Funded by the Austrian Science Fund (FWF KLI 1012) and the Vienna Science and Technology Fund (WWTF LS22-056). No funder influenced study design, data collection, analysis, or publication.

Conflict of Interest

The author declares no competing interests and no financial relationships with AI diagnostics vendors or emergency medicine technology providers.

Data Availability Statement

Aggregated ED-level outcome data and the TASS calculation tool are available from the author. Patient-level data access requires institutional ethics committee approval from each participating hospital.

Ethical Approval

Approved by Vienna General Hospital Ethics Committee (EK-2019-1042), Gemelli Hospital Rome (Prot. 2019-0218), Karolinska Regional Ethics Stockholm (Dnr 2019-03844), and CHLN Ethics Committee Lisbon (CHLN-2019-312).

Appendix A

TASS Scoring Rubric and AI System Technical Specifications

Full scoring criteria for each TASS dimension and technical specifications of each AI triage system architecture.