International Journal of Innovative Technologies in Social Science

e-ISSN: 2544-9435

Operating Publisher
SciFormat Publishing Inc.
ISNI: 0000 0005 1449 8214

2734 17 Avenue SW,
Calgary, Alberta, T3E0A7,
Canada
+15878858911
editorial-office@sciformat.ca

ARTICLE TITLE	ARTIFICIAL INTELLIGENCE IN EMERGENCY DEPARTMENT TRIAGE: A NARRATIVE REVIEW OF ALGORITHMIC ACCURACY, RACIAL AND SOCIOECONOMIC DISPARITIES, AND THE INTEGRATION OF SOCIAL DETERMINANTS OF HEALTH
DOI	https://doi.org/10.31435/ijitss.2(50).2026.5675
RECEIVED	26 March 2026
ACCEPTED	24 June 2026
PUBLISHED	29 June 2026
LICENSE	The article is licensed under a **Creative Commons Attribution 4.0 International License.**

© The author(s) 2026.
This article is published as open access under the Creative Commons Attribution 4.0 International License (CC BY 4.0), allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

ARTIFICIAL INTELLIGENCE IN EMERGENCY DEPARTMENT TRIAGE: A NARRATIVE REVIEW OF ALGORITHMIC ACCURACY, RACIAL AND SOCIOECONOMIC DISPARITIES, AND THE INTEGRATION OF SOCIAL DETERMINANTS OF HEALTH

Damian Danilczuk *(Corresponding Author, Email: danilczukdamian96@gmail.com)*
John Paul II Municipal Hospital in Rzeszów, Rycerska 4, 35-241 Rzeszów, Poland
ORCID ID: 0009-0005-5755-1115

Jerzy Buszko
Medical Care Center in Jarosław, 3-go Maja 70, 37-500 Jarosław, Poland
ORCID ID: 0009-0008-2708-1222

Michał Dyś
Health Care Complex in Dębica, Krakowska 91, 39-200 Dębica, Poland
ORCID ID: 0009-0000-2644-7626

Nicol Baran
Medical Center in Łańcut, Ignacego Paderewskiego 5, 37-100 Łańcut, Poland
ORCID ID: 0009-0003-3607-6744

Marta Bajkowska-Piterak
Medical Center in Łańcut, Ignacego Paderewskiego 5, 37-100 Łańcut, Poland
ORCID ID: 0009-0000-6689-0788

Przemysław Piterak
Hospital of the Ministry of Interior and Administration in Rzeszów, Krakowska 16, 35-111 Rzeszów, Poland
ORCID ID: 0009-0002-3553-7353

Monika Szlachta-Gubernat
Hospital of the Ministry of Interior and Administration in Rzeszów, Krakowska 16, 35-111 Rzeszów, Poland
ORCID ID: 0009-0005-6994-9761

Wiktor Adamiec
John Paul II Municipal Hospital in Rzeszów, Rycerska 4, 35-241 Rzeszów, Poland
ORCID ID: 0009-0002-3322-5479

Adrianna Buż
John Paul II Municipal Hospital in Rzeszów, Rycerska 4, 35-241 Rzeszów, Poland
ORCID ID: 0009-0001-8202-5628

ABSTRACT

Background: Emergency department (ED) crowding represents a growing global public health challenge. Limitations of conventional five-level triage scales — including interrater variability and documented racial and socioeconomic disparities — have driven increasing interest in artificial intelligence (AI) for triage optimization. Existing literature has largely examined algorithmic accuracy, fairness, and social determinants of health (SDOH) in isolation.
Objective: This narrative review synthesizes evidence across three interconnected domains: algorithmic accuracy of AI triage models compared with conventional scales; racial and socioeconomic disparities and the mechanisms producing algorithmic bias; and the integration of SDOH into AI-driven triage.
Methods: Peer-reviewed publications were retrieved from PubMed/MEDLINE, Embase, Scopus, Web of Science, and IEEE Xplore (January 2015 through March 2026), combining structured keyword-based searches across emergency care, AI methodology, equity, and SDOH.
Results: AI models — particularly gradient-boosted ensembles and NLP-augmented architectures — consistently outperform ESI, MTS, and CTAS in predicting critical endpoints, with AUROC gains of 0.05–0.10. However, biased proxy variables, measurement artifacts (e.g., pulse oximetry), unrepresentative training data, and biased outcome labels can amplify existing disparities. SDOH integration offers potential for improving equity but faces challenges including Z-code underutilization, definitional heterogeneity, and stigmatization risk. Emerging frameworks — the EU AI Act, FDA guidance, WHO principles, and TRIPOD+AI — provide a scaffold for responsible deployment.
Conclusions: Realizing AI's benefits while avoiding harm requires rigorous external validation, transparent subgroup reporting, explicit fairness evaluation, post-deployment monitoring, and participatory design.

KEYWORDS

Artificial Intelligence; Emergency Triage; Algorithmic Bias; Health Equity; Social Determinants of Health

CITATION

Damian Danilczuk, Jerzy Buszko, Michał Dyś, Nicol Baran, Marta Bajkowska-Piterak, Przemysław Piterak, Monika Szlachta-Gubernat, Wiktor Adamiec, Adrianna Buż. (2026) Artificial Intelligence in Emergency Department Triage: a Narrative Review of Algorithmic Accuracy, Racial and Socioeconomic Disparities, and the Integration of Social Determinants of Health. *International Journal of Innovative Technologies in Social Science*. 2(50). doi: 10.31435/ijitss.2(50).2026.5675

COPYRIGHT

© The author(s) 2026. This article is published as open access under the **Creative Commons Attribution 4.0 International License (CC BY 4.0)**, allowing the author to retain copyright. The CC BY 4.0 License permits the content to be copied, adapted, displayed, distributed, republished, or reused for any purpose, including adaptation and commercial use, as long as proper attribution is provided.

1. Introduction

Emergency departments (EDs) serve as a critical gateway to acute healthcare systems worldwide. Over the past two decades, ED visit volumes have risen substantially, with the United States alone recording approximately 140 million annual visits in 2021, corresponding to a rate of 43 visits per 100 persons (Cairns et al., 2023). This trend is not confined to high-income settings; global analyses indicate that emergency conditions account for more than half of all deaths globally, yet emergency care utilization rates remain disproportionately low in low- and middle-income countries, where the burden of acute disease is highest (C. Y. Chang et al., 2016). The resulting mismatch between rising demand and constrained capacity has made ED crowding — defined by the American College of Emergency Physicians as a situation in which the identified need for emergency services exceeds available resources — a pervasive global public health crisis (Pearce et al., 2023). Systematic reviews have consistently demonstrated that crowding is associated with delayed treatment initiation, increased short-term mortality, elevated rates of patients leaving without being seen, and diminished quality of care (Morley et al., 2018; Rasouli et al., 2019).

Triage — the process of rapidly prioritizing patients according to clinical urgency — constitutes the principal mechanism through which EDs manage patient flow under conditions of demand–capacity mismatch. Several structured five-level triage scales have been adopted internationally, most notably the Emergency Severity Index (ESI), the Manchester Triage System (MTS), and the Canadian Triage and Acuity Scale (CTAS). While these instruments have been widely validated and represent significant advances over earlier, less standardized approaches, they carry well-documented limitations (Zachariasse et al., 2019). Conventional triage relies heavily on the clinical judgment of triage nurses, introducing subjectivity, interrater variability, and susceptibility to cognitive and implicit biases. A large multicenter cohort study has further revealed that

mistriage under ESI is not uniformly distributed: Black patients experienced a significantly greater relative risk of both undertriage and overtriage compared with White patients, and patients from lower-income neighborhoods were also disproportionately affected (Sax et al., 2023; see Section 3.1 for full performance figures). These findings underscore that the limitations of conventional triage are not merely operational but carry equity implications.

Against this background, artificial intelligence (AI) and machine learning (ML) have emerged as promising tools to augment or potentially transform ED triage. The application of AI in this domain has evolved considerably, progressing from early rule-based clinical decision support systems to sophisticated approaches involving gradient-boosted decision trees, deep neural networks, natural language processing (NLP) of clinical free-text, and, more recently, multimodal models integrating structured and unstructured data (Tyler et al., 2024; Porto, 2024). Multiple systematic reviews have demonstrated that ML models — particularly ensemble methods such as XGBoost and gradient boosting — consistently outperform conventional triage scales in predicting critical clinical endpoints, including in-hospital mortality, intensive care unit (ICU) admission, and the need for life-saving interventions (Almulihi et al., 2024; El Arab & Al Moosa, 2025). Furthermore, NLP techniques have been shown to enhance classification performance when applied to triage nursing notes, enabling algorithms to capture nuances in clinical presentation that structured data alone may miss (Porto, 2024).

However, the growing enthusiasm for AI-driven triage must be tempered by critical examination of potential harms. A foundational study by Obermeyer et al. (2019), published in *Science*, revealed that a widely used commercial algorithm for identifying high-risk patients exhibited significant racial bias — at any given risk score, Black patients were substantially sicker than White patients, because the algorithm used healthcare costs as a proxy for health needs, thereby encoding historical disparities in access to care. Although this study did not focus specifically on ED triage, it catalyzed a broader reckoning with algorithmic bias in clinical AI and established a conceptual framework — the problem of biased proxy variables — that is directly applicable to triage contexts (Vyas et al., 2020; Gianfrancesco et al., 2018). Subsequent research has documented that AI models trained on data from predominantly White, urban, and well-insured populations may perform poorly when deployed in settings serving racial and ethnic minorities, rural communities, or socioeconomically disadvantaged groups (Rajkomar et al., 2018; Chen et al., 2023).

An additional dimension, largely underexplored in the AI triage literature, concerns the integration of social determinants of health (SDOH) — the conditions in which people are born, grow, live, work, and age that shape health outcomes (World Health Organization, 2010). Medical care contributes to only approximately 20% of modifiable health outcomes, with the remainder influenced by social and behavioral factors (Hood et al., 2016). Recent scoping reviews have identified a growing body of work leveraging NLP and ML techniques to extract SDOH data from clinical notes in ED settings, yet this convergence remains in its infancy, with significant challenges related to data completeness, definitional heterogeneity, and the risk of algorithmic stigmatization of already marginalized populations (Abbott et al., 2024).

Despite the rapid expansion of the literature on AI in emergency triage, existing reviews have tended to address these issues in isolation — focusing either on algorithmic performance, on bias and fairness, or on social determinants — without providing an integrated analysis of their interrelationships. This fragmentation limits the ability of clinicians, policymakers, and algorithm developers to appreciate the full spectrum of considerations necessary for the equitable deployment of AI-assisted triage systems.

The present narrative review aims to address this gap by synthesizing evidence across three interconnected domains: (1) the algorithmic accuracy of AI models for ED triage compared with conventional triage scales; (2) racial and socioeconomic disparities in algorithmic performance and the mechanisms through which bias arises; and (3) the potential and challenges of incorporating SDOH into triage algorithms. In doing so, the review seeks to answer the following research questions:

1. How do AI-based triage models compare with conventional triage systems in predicting critical clinical outcomes, and what are the key limitations of current evidence regarding their generalizability?
2. What are the documented racial and socioeconomic disparities in AI triage performance, and through what mechanisms do algorithmic biases emerge?
3. To what extent have SDOH been integrated into AI-driven triage models, and what are the benefits, risks, and implementation challenges of such integration?
4. What ethical frameworks, regulatory instruments, and bias mitigation strategies are available to guide the equitable development and deployment of AI in emergency triage?

By addressing these questions within a unified analytical framework, this review intends to provide actionable insights for researchers designing equitable algorithms, clinicians evaluating AI tools for clinical adoption, and policymakers developing regulatory standards for AI in emergency medicine.

2. Methodology

2.1. Rationale for a Narrative Review Approach

A narrative review methodology was selected as the most appropriate approach to address the multifaceted research objectives of this study. Narrative reviews are particularly well-suited to the synthesis of heterogeneous bodies of literature spanning distinct methodological traditions, including clinical validation studies of artificial intelligence (AI) models, empirical investigations of health disparities, analyses of algorithmic bias, and conceptual frameworks related to social determinants of health (SDOH) (Sukhera, 2022; Ferrari, 2015).

Unlike systematic reviews, which are typically designed to address narrowly defined research questions amenable to quantitative synthesis, narrative approaches enable the integration of conceptually diverse evidence, the identification of cross-disciplinary patterns, and the development of interpretative frameworks that extend beyond individual study designs (Greenhalgh et al., 2018; Grant & Booth, 2009).

In this study, the narrative approach supports the development of an integrative analytical framework linking three interrelated domains: algorithmic accuracy, bias and disparities, and the incorporation of SDOH into AI-driven clinical decision-making. Methodological rigor was guided by the principles outlined in the Scale for the Assessment of Narrative Review Articles (SANRA), including clarity of objectives, transparency in literature selection, and structured presentation of evidence (Baethge et al., 2019).

2.2. Literature Search Strategy

A targeted literature search was conducted to identify relevant peer-reviewed publications across clinical, public health, and technical domains. Major biomedical and interdisciplinary databases, including PubMed/MEDLINE, Scopus, and Web of Science, were consulted. In addition, IEEE Xplore was included to capture technical and engineering-oriented contributions to AI model development that may not be fully represented in biomedical indexing systems.

Search terms combined four conceptual domains:

1) emergency care and triage,
2) artificial intelligence and machine learning,
3) algorithmic bias and health equity, and
4) social determinants of health.

The search focused primarily on literature published between 2015 and 2026, reflecting the period of rapid expansion in clinical AI applications and increasing scholarly attention to fairness and equity. Seminal earlier works were included where they contributed foundational concepts (e.g., Obermeyer et al., 2019; Gianfrancesco et al., 2018).

Titles and abstracts were screened for relevance, followed by full-text assessment of selected articles. Rather than aiming for exhaustive retrieval, a purposive sampling strategy was employed to identify studies offering high conceptual relevance and methodological diversity. Priority was given to high-impact empirical studies, systematic and scoping reviews, and influential conceptual or policy-oriented publications.

2.3. Analytical Approach

The included literature was analyzed using an interpretative, thematic synthesis approach. Studies were organized into three core analytical domains:

1) algorithmic accuracy of AI-based triage models,
2) mechanisms of racial and socioeconomic bias in clinical prediction systems, and
3) integration of social determinants of health into algorithmic design and implementation.

These domains served as an organizing framework for cross-domain comparison and conceptual integration. The synthesis focused on identifying recurring themes, methodological tensions, and structural gaps in the current evidence base, particularly at the intersection of technical model performance and health equity considerations.

Consistent with the objectives of a narrative review, no formal risk-of-bias assessment or meta-analysis was conducted. The aim was to provide critical interpretation and conceptual integration rather than quantitative effect estimation.

2.4. Methodological Limitations

Several limitations inherent to the narrative review methodology should be acknowledged. First, the absence of a formal systematic protocol and risk-of-bias assessment limits the ability to evaluate the strength of evidence across studies. Second, the purposive selection of literature introduces potential selection bias, as inclusion is influenced by author judgment.

Third, publication bias—particularly the preferential reporting of studies demonstrating favorable algorithmic performance—may affect the available evidence base. Fourth, the rapid evolution of AI research means that the findings of this review represent a time-bounded synthesis that may be subject to change as new evidence emerges.

Finally, the interpretative nature of thematic synthesis introduces an additional layer of subjectivity in the identification, categorization, and framing of key themes and relationships across domains.

3. Traditional Triage Systems and AI Approaches — Conceptual Foundations

3.1. Conventional Five-Level Triage Scales and Their Limitations

Modern ED triage is dominated by structured five-level scales that replaced earlier, unstandardized sorting methods. Three systems have achieved the greatest international penetration and together shape most contemporary debates about triage accuracy: the Emergency Severity Index (ESI), used in more than 70% of US EDs; the Manchester Triage System (MTS), dominant across the United Kingdom and continental Europe; and the Canadian Triage and Acuity Scale (CTAS), the standard across Canada and increasingly adopted in the Middle East and Asia (Zachariasse et al., 2019). All three stratify patients into five priority categories with recommended maximum times to physician evaluation, from immediate (Level I) to non-urgent (Level V).

Despite shared structure, the scales differ substantially in their operating logic. ESI uses a branching algorithm in which the nurse first assesses immediate life-threatening criteria, then anticipated resource utilization, incorporating vital signs only as tiebreakers at mid-acuity levels. MTS employs approximately 50 flowcharts organized around presenting complaints (e.g., "chest pain," "abdominal pain in an adult"), with discriminators progressively narrowing the acuity category (Gräff et al., 2014). CTAS, updated most recently in 2025, relies on an ordered list of 165 presenting complaints combined with first-order modifiers (vital signs, pain scales, mechanism of injury) and second-order modifiers specific to the complaint group (Hall et al., 2025; Bullard et al., 2017).

Empirical evidence regarding the performance of these instruments is mixed. A meta-analysis of 14 studies encompassing heterogeneous settings reported that MTS demonstrated substantial overall interrater reliability, with pooled weighted kappa values around 0.75, although agreement was notably lower in mental-health presentations and at the margins of the urgency spectrum (Mirhaghi et al., 2017). CTAS has shown comparable reliability, with kappa coefficients of 0.48 (unweighted) and 0.71 (weighted), and performance improving substantially with structured training and use of second-order modifiers (Mirhaghi et al., 2015). For ESI, the evidence is more sobering: the large multicenter study by Sax et al. (2023) demonstrated that roughly one in three ED encounters were mistriaged, with sensitivity of only 65.9% for identifying critically ill patients. A recurrent finding across instruments is that subjectivity in nurse judgment, interrater variability, susceptibility to cognitive and implicit biases, and limited ability to incorporate complex clinical reasoning produce systematic patterns of under- and over-triage (Hinson et al., 2018; Zachariasse et al., 2019). Furthermore, the relatively crude five-level granularity compresses substantial within-level heterogeneity, with more than 60% of encounters typically funneled into the middle category, thereby limiting the scale's discriminative utility (Sax et al., 2023).

3.2. A Typology of AI Methods Applied to Emergency Triage

A wide range of computational methods has been applied to ED triage, broadly spanning four categories: classical machine learning, deep learning, natural language processing, and multimodal approaches.

Classical machine learning methods — including logistic regression, random forests, support vector machines, and gradient-boosted decision trees such as XGBoost — have dominated the early wave of triage AI research. Their comparative advantages are interpretability, modest data requirements, and computational efficiency. Ensemble tree-based models, in particular, have consistently demonstrated strong performance in triage prediction tasks; a recent systematic review of 17 studies concluded that boosting algorithms most often outperformed both competing ML approaches and conventional triage scales in predicting admission, ICU transfer, and critical intervention need (Almulihi et al., 2024). For example, Yun et al. (2021) trained an XGBoost classifier on 13 routinely collected triage variables from a Korean academic ED (n = 80,433) and

achieved superior discrimination over a baseline logistic model using the Korean Triage and Acuity Scale (KTAS).

Deep learning approaches — feed-forward and recurrent neural networks, and more recently transformer-based architectures — have extended the state of the art, particularly for tasks involving large, heterogeneous data or unstructured clinical text. Grant et al. (2025), using retrospective data from 670,841 CTAS-triaged visits, trained three models (LASSO regression, gradient-boosted trees, and a deep learning network with embeddings) and reported AUROC values of 0.89–0.93 for predicting the need for critical care within 12 hours, compared with 0.80 for CTAS alone. Deep learning offers two principal advantages: the capacity to learn hierarchical feature representations directly from raw data, and the ability to integrate temporal dynamics (e.g., trends in vital signs) through recurrent layers.

Natural language processing (NLP) has emerged as a distinct and rapidly growing methodological stream, reflecting the recognition that triage nursing notes and chief complaints contain clinically rich information not captured in structured fields. Porto (2024), in a systematic review, concluded that NLP-augmented triage models consistently outperformed their structured-data-only counterparts. Techniques range from classical approaches such as term-frequency/inverse-document-frequency vectors, through pretrained embeddings, to clinically adapted large language models. Multilingual sentence embeddings applied to free-text chief complaints contributed substantially to the XGBoost model developed by Sitthiprawiat et al. (2025), which achieved AUROC 0.92 for ICU admission prediction in a Thai academic center (n = 163,452).

Finally, multimodal models integrate structured tabular data, free-text notes, and, increasingly, physiological waveform or imaging data. Although still uncommon in operational deployment, such approaches are at the frontier of triage AI, offering the potential to capture the full clinical Gestalt that skilled clinicians intuitively synthesize at the bedside (Tyler et al., 2024).

3.3. Typology of Input Data

The performance of any AI triage model is critically contingent on the quality and composition of its input data. Four broad categories predominate in the literature. First, demographic variables — age, sex, race/ethnicity, mode of arrival, and insurance status — are ubiquitous but their use has equity implications, discussed in Section 5. Second, vital signs collected at presentation (blood pressure, heart rate, respiratory rate, oxygen saturation, temperature, level of consciousness) constitute the single most important predictive block across studies. Third, information extracted from the electronic health record (EHR), including prior diagnoses, medications, prior healthcare utilization, and comorbidity indices, substantially improves model performance: Hong et al. (2018), analyzing 560,486 visits, demonstrated that adding EHR-derived history to triage variables increased AUROC for hospital admission from 0.87 to 0.92. Fourth, laboratory and point-of-care results, though typically unavailable at the moment of triage, may be incorporated in dynamic models that update predictions as results become available during the ED stay. An emerging fifth category comprises unstructured clinical text — chief complaints and triage nursing notes — whose incorporation via NLP has become a defining feature of state-of-the-art models (Porto, 2024).

The choice of input variables is not a technical decision alone; it carries profound implications for model generalizability, fairness, and interpretability, themes developed in the sections that follow.

4. Algorithmic Accuracy of AI Triage Models

4.1. Evaluation Metrics

Rigorous evaluation of AI triage models requires a suite of complementary performance metrics, each capturing a different facet of model utility. Discrimination — the ability to rank patients correctly by risk — is typically quantified by the area under the receiver operating characteristic curve (AUROC), with values above 0.80 considered acceptable and above 0.90 excellent for binary outcomes. However, AUROC can be misleading in the highly imbalanced settings characteristic of ED triage, where critical endpoints (e.g., ICU admission, in-hospital mortality) affect fewer than 10% of patients. In such cases, the area under the precision–recall curve (AUPRC) provides a more informative assessment of a model's ability to identify the minority positive class (Sitthiprawiat et al., 2025). Complementary metrics include sensitivity and specificity at clinically meaningful thresholds, positive and negative predictive values, and the F1 score. Calibration — the agreement between predicted probabilities and observed event rates — is essential for clinical decision-making but is frequently overlooked in the AI triage literature; a well-discriminating but poorly calibrated model can systematically over- or underestimate risk, leading to misallocation of resources (Yun et al., 2021). The TRIPOD+AI guideline, published in 2024, formalizes these expectations, mandating transparent reporting of discrimination, calibration, subgroup performance, and uncertainty quantification for all clinical prediction models (Collins et al., 2024).

4.2. Prediction of Critical Clinical Endpoints

AI triage models have been evaluated against a broad spectrum of clinically meaningful endpoints. The most frequently studied outcomes include hospital admission, ICU admission, critical care need (often defined as the composite of ICU admission and in-hospital mortality), the need for specific life-saving interventions, and short-term mortality. For admission prediction, ML models consistently achieve AUROC values in the 0.85–0.92 range, substantially exceeding those of conventional five-level scales (Hong et al., 2018; Almulihi et al., 2024). For critical care outcomes, Yun et al. (2021) reported an XGBoost AUROC of 0.86 compared with 0.80 for KTAS-based logistic regression in a Korean ED cohort, while Sitthiprawiat et al. (2025) reported AUROC 0.92 for their XGBoost model versus 0.88 for CTAS in a Thai cohort of 163,452 visits. Beyond composite endpoints, H. Chang et al. (2022) developed an XGBoost model predicting the need for six specific critical interventions — arterial line placement, oxygen therapy, high-flow nasal cannula, intubation, massive transfusion, and vasopressor administration — with AUROCs ranging from 0.91 to 0.96 using only triage-available variables. These results suggest that AI can move beyond global severity ranking to enable more granular, action-oriented risk stratification at the point of triage.

4.3. Comparative Performance Against Conventional Triage Scales

A consistent finding across systematic reviews is the superior discriminative performance of ML approaches compared with conventional triage scales. Tyler et al. (2024), in a scoping review of 29 studies, found that ML models consistently demonstrated better discrimination than ESI, MTS, or CTAS in predicting downstream clinical outcomes. Almulihi et al. (2024), synthesizing 17 studies, reported that boosting algorithms were generally the most accurate, with gradient boosting and XGBoost outperforming logistic regression, decision trees, and traditional scales. The systematic review by Porto (2024), focused specifically on NLP-enhanced triage, concluded that incorporating triage nurse notes and chief-complaint free text via NLP consistently improved classification performance compared with algorithms using only structured variables. Grant et al. (2025) provided perhaps the most compelling head-to-head comparison: using 670,841 ED visits, three ML models achieved AUROCs of 0.89–0.93 for early critical care need, versus 0.80 for CTAS alone, with the corresponding AUPRC more than doubling (0.23–0.27 vs. 0.11).

The evidence base has begun to move beyond retrospective benchmark comparisons to prospective implementation trials. Taylor et al. (2025), in a multisite prospective evaluation published in *NEJM AI*, reported that an AI-informed triage clinical decision support (CDS) system increased sensitivity for detecting critical illness from 78.8% to 83.1% and reduced time to disposition decision. A companion analysis from the same platform also documented reductions in undertriage disparities across racial and ethnic groups (Hinson, Levin, et al., 2025); this equity-related finding is discussed in detail in Section 5.2.

4.4. External Validation and Generalizability

Despite the impressive performance of AI models in development cohorts, the generalizability of these models across institutions, geographic regions, and patient populations remains a critical concern. Models trained on data from a single academic center frequently exhibit performance degradation when validated externally due to differences in patient case-mix, documentation practices, coding conventions, and workflow. This phenomenon — known as dataset shift — has prompted increased emphasis on rigorous external validation prior to deployment. The integrative systematic review by El Arab and Al Moosa (2025) concluded that the overwhelming majority of published ED triage AI studies report only internal validation, with fewer than a quarter providing any form of external or multicenter evaluation. This gap is particularly concerning given evidence that models can exhibit substantial performance deterioration and fairness drift when applied to populations that differ from the training distribution.

Overfitting constitutes a related methodological pitfall. When models are trained and evaluated on data from the same institution without appropriate regularization, cross-validation, and calibration strategies, reported performance may substantially overestimate the expected real-world utility. The TRIPOD+AI checklist explicitly addresses these concerns, requiring authors to report training–test splits, internal and external validation procedures, sample-size justifications, and performance metrics with confidence intervals (Collins et al., 2024). Nevertheless, audits of the published ED triage AI literature suggest that adherence to these reporting standards remains inconsistent, limiting reproducibility and meaningful cross-study comparisons.

Finally, the temporal stability of model performance — the phenomenon of "model drift" as underlying patient populations, clinical practices, or documentation conventions evolve — is an increasingly recognized concern. Few published ED triage AI studies report on post-deployment monitoring or incorporate mechanisms for continuous performance surveillance. The US Food and Drug Administration's guidance on Predetermined

Change Control Plans for adaptive ML-enabled devices attempts to address this issue by requiring manufacturers to specify how models will be updated, retrained, and monitored after market authorization, but operationalizing such plans in hospital settings remains nascent.

In summary, the available evidence supports a cautiously optimistic conclusion: AI models, and particularly tree-based ensembles and NLP-augmented approaches, can achieve meaningfully better discrimination and risk stratification than conventional five-level scales in the prediction of critical ED endpoints. However, the translation of these technical gains into clinical benefit remains contingent on rigorous external validation, transparent reporting, attention to calibration and subgroup performance, and sustained post-deployment monitoring — a challenging agenda that the field is only beginning to address.

5. Algorithmic Bias: Racial and Socioeconomic Dimensions

5.1. A Taxonomy of Algorithmic Bias

Algorithmic bias in clinical AI does not emerge from malicious intent but from the complex interplay of data, modeling choices, and the sociotechnical context of deployment. Several interrelated but conceptually distinct mechanisms can produce biased outcomes. *Data bias* arises when training datasets fail to adequately represent the populations in which the model will ultimately be deployed — for example, when minority racial groups, rural patients, or non-English speakers are systematically underrepresented in academic EHR repositories (Gianfrancesco et al., 2018). *Label bias* occurs when the outcome variable itself reflects historical inequities: if a model is trained to predict who "received" intensive treatment rather than who "needed" it, it will encode past disparities in access (Obermeyer et al., 2019). *Measurement bias* arises when clinical measurements themselves differ systematically across groups, as with pulse oximetry readings in patients with darker skin (Sjoding et al., 2020). *Representation bias* in embeddings — particularly in NLP models trained on biased medical text — can propagate implicit stereotypes into algorithmic predictions. Finally, *deployment bias* occurs when a model is used in ways or settings that differ from its intended use, or when human operators apply its outputs selectively across patient groups.

5.2. Mechanisms of Racial Disparity in AI Triage

The foundational study by Obermeyer et al. (2019), while not focused on ED triage, provides the clearest empirical demonstration of how proxy-variable bias can produce clinically significant racial disparities. The authors analyzed a commercial algorithm used to identify patients for enrollment in complex care management programs, affecting millions of Americans. Because the algorithm predicted healthcare costs rather than health needs, and because Black patients incurred lower costs at any given level of illness due to structural barriers to care, the algorithm systematically underestimated Black patients' needs. Correcting the bias would have increased the proportion of Black patients automatically enrolled from 17.7% to 46.5%. The mechanism — reliance on a cost proxy that encoded unequal access — is directly applicable to ED triage models that use prior utilization or spending as features.

A second, particularly concerning mechanism is measurement bias in physiological signals themselves. Sjoding et al. (2020), analyzing more than 10,000 paired measurements of pulse oximetry and arterial blood gas saturation, found that Black patients had nearly three times the rate of occult hypoxemia (arterial saturation below 88% despite pulse oximetry readings of 92–96%) compared with White patients. This bias has substantial implications for AI triage models that incorporate oxygen saturation — a near-universal input — because the same numerical SpO_2 value corresponds to different true physiological states across racial groups. Any model that treats pulse oximetry as race-neutral will propagate this bias into its risk estimates, systematically underestimating severity in Black patients.

A third mechanism concerns race-based clinical calculators. Vyas et al. (2020) documented the use of race correction factors in numerous widely deployed clinical algorithms — from eGFR and pulmonary function predictions to obstetric risk scores — arguing that many such corrections are empirically unsupported and propagate racist assumptions about biological difference. The implications for triage are twofold: first, AI models may inherit these biased base rates if they use race-adjusted outcome labels as training targets; second, models may "learn" to encode race implicitly through correlated variables (e.g., ZIP code, insurance type) even when race is explicitly excluded as a predictor.

Empirical evidence specific to ED triage bias is emerging. Sax et al. (2023) demonstrated that conventional ESI itself exhibits racial disparities, with Black patients experiencing an 18.5% greater relative risk of undertriage than White patients. When AI models are trained on ESI-labeled data, they risk inheriting and amplifying this bias. Conversely, appropriately designed AI tools may offer corrective potential: Hinson, Levin, et al. (2025), evaluating a multisite implementation of AI-assisted triage, reported reductions in

undertriage disparities across racial and ethnic groups, suggesting that the encoding of a richer, more objective feature set may offset some of the implicit biases that affect nurse judgment.

5.3. The Socioeconomic Dimension

Racial disparities in algorithmic triage are inextricably intertwined with socioeconomic inequality. Operationalizing socioeconomic status (SES) in AI research is non-trivial: individual-level measures (income, education, insurance status) are often incomplete in EHR data, prompting reliance on area-level proxies such as ZIP code, neighborhood deprivation indices, or area deprivation scores. Each approach carries trade-offs. Insurance status, while widely available, reflects not only SES but also legal status and employment type; area-level proxies risk ecological fallacy when individual-level variation is high.

Beyond measurement, the underrepresentation of low-SES populations in training datasets is a consistent concern. Large ED AI studies have disproportionately drawn on data from academic tertiary centers serving urban, insured populations, with limited representation from safety-net hospitals, rural EDs, and critical access hospitals. This distribution is a form of selection bias that may produce models that perform poorly precisely in the settings where they are most needed. Geographic inequalities compound these effects: rural EDs often have smaller patient volumes, different case-mix, and more limited laboratory capacity, all of which can degrade the performance of models trained in high-resource urban settings (Chen et al., 2023).

An additional and under-recognized mechanism concerns insurance-related proxy variables. AI models that include insurance type as a feature — directly or indirectly — may encode the care access gap as a medical finding, perpetuating the kind of proxy bias identified by Obermeyer et al. (2019). The problem is particularly acute when models are trained on outcomes influenced by resource allocation (e.g., hospitalization decisions), which are themselves shaped by insurance coverage.

5.4. Intersectionality and Compounded Disadvantage

A critical limitation of much existing fairness research is its focus on single protected attributes — race, sex, or SES — rather than their intersections. In reality, patients experience compounded disadvantages: Sax et al. (2023) reported that Black male patients in EDs had a 41.0% greater relative risk of undertriage compared with White female patients, substantially larger than the effect of race alone. These intersectional effects can be masked by aggregate fairness metrics and require explicit stratified evaluation across combinations of protected attributes. Yet few published AI triage studies report subgroup performance at this level of granularity.

5.5. Fairness Metrics and Their Trade-offs

Operationalizing fairness in AI requires translating intuitive concepts of equity into mathematical criteria. Three group-fairness metrics have become standard (Chen et al., 2023). *Demographic parity* (or statistical parity) requires that the probability of a positive prediction be equal across protected groups; its limitation in healthcare is that it can be inappropriate when true base rates of disease genuinely differ between groups. *Equalized odds* requires that both true-positive and false-positive rates be equal across groups, conditional on the true outcome; it does not enforce equal prediction rates but does require comparable classification quality. *Calibration within groups* requires that predicted probabilities correspond to observed event rates uniformly across groups. Critically, the seminal work of Kleinberg et al. (2017) has shown that these criteria cannot, in general, be satisfied simultaneously except in degenerate cases — a mathematical impossibility theorem with profound practical implications. The selection among fairness criteria therefore reflects a value judgment about which form of equity matters most in a given clinical context, not merely a technical optimization.

Fairness metrics themselves presuppose that ground-truth labels are unbiased — an assumption that is frequently violated in healthcare, where outcome labels such as “admitted” or “received intervention” reflect historical inequities in the care pathway. Conditioning on such labels, as equalized odds does, may entrench rather than correct existing disparities. This observation motivates an important caution: fairness metrics are necessary but not sufficient instruments for equity, and their deployment must be accompanied by substantive attention to the origin and validity of the labels on which models are trained.

6. Integration of Social Determinants of Health into Triage Algorithms

6.1. Conceptual Frameworks

Social determinants of health (SDOH) — the conditions in which people are born, grow, live, work, and age, and the broader forces shaping those conditions — are now widely recognized as the dominant drivers of population health, accounting for 30%–55% of health outcomes according to WHO estimates, and far exceeding the contribution of clinical care (Hood et al., 2016; World Health Organization, 2010). The WHO's *Conceptual framework for action on the social determinants of health* and the US *Healthy People 2030* initiative both organize SDOH into domains including economic stability, education access and quality, healthcare access and quality, neighborhood and built environment, and social and community context. In the ED, SDOH shape both the probability of presentation and the nature of clinical trajectories: housing insecurity, food insecurity, and transportation barriers are independently associated with higher ED utilization, hospitalization risk, and post-discharge adverse events. The integration of SDOH into AI triage thus holds theoretical promise to improve both predictive accuracy and equity of care, but translates into substantial data, methodological, and ethical challenges.

6.2. Data Sources for SDOH in the Emergency Department

Four principal data sources enable SDOH capture in ED AI pipelines. The first is the set of ICD-10-CM Z-codes (categories Z55–Z65), which include standardized codes for problems related to education, employment, housing, the social environment, and economic circumstances. Although Z-codes represent a potentially powerful structured data source, they remain severely underutilized: analyses of the US National Inpatient Sample for 2016–2017 found SDOH Z-codes in only 1.9% of admissions, with similar underreporting persisting in outpatient and ED settings (Truong et al., 2020). Barriers to adoption include the absence of reimbursement incentives, workflow constraints, ambiguity about which clinicians may document Z-codes, and limited training in their use.

The second source is natural language processing (NLP) applied to unstructured clinical text. Triage nursing notes, history and physical examinations, and social work consultations frequently contain references to SDOH that are not captured in structured fields. Abbott et al. (2024), in a scoping review of 26 studies, documented growing use of rule-based, classical ML, and deep learning NLP techniques to extract SDOH from clinical notes in ED and emergency-medicine contexts, with F1 scores ranging from 0.40 to 0.75 across predicted SDOH-related outcomes, with homelessness and neighborhood or built-environment factors among the most frequently studied domains. However, the review also noted significant variability in performance across domains, with rarer or more nuanced SDOH (e.g., interpersonal violence, literacy) proving more difficult to extract reliably.

The third source is patient-reported screening tools administered in clinical workflows. The Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE) — a 20-item validated instrument developed by the National Association of Community Health Centers — and the Accountable Health Communities Health-Related Social Needs (AHC-HRSN) screening tool from the US Centers for Medicare and Medicaid Services are the most widely adopted instruments. Both have been integrated into EHRs at some academic medical centers, enabling structured capture of SDOH as input features for predictive models (Moen et al., 2020). Real-world deployment, however, reveals substantial missingness and non-response; ED patients may be reluctant to complete lengthy instruments during acute presentations, particularly when the workflow does not allow for immediate response to identified needs.

The fourth source is area-level geospatial data, including neighborhood-level socioeconomic indices (e.g., the Area Deprivation Index, the Social Vulnerability Index) linked to patient address at the time of presentation. Such linkage allows inference of contextual exposures even when individual-level SDOH data are missing, though it carries the risk of ecological fallacy.

6.3. Implementation Challenges

The operational integration of SDOH into AI triage models faces several interconnected challenges. Definitional heterogeneity is pervasive: different instruments define the same construct (e.g., food insecurity) differently, and mapping between them is imperfect. Data completeness is a persistent limitation; as Abbott et al. (2024) emphasize, SDOH data in EHRs are typically missing not at random but rather associated with patient characteristics that independently influence outcomes, creating a form of informative missingness that complicates modeling. The time-intensive nature of SDOH assessment conflicts with ED workflow demands, particularly during crowding. Data standardization efforts — notably the Gravity Project's FHIR SDOH Clinical Care Implementation Guide — aim to address definitional heterogeneity but have not yet achieved widespread adoption.

A further concern is the risk of stigmatization and privacy infringement. Flagging patients as high-risk on the basis of homelessness, immigration status, or substance use disorder can produce real harms — diverted care, altered clinical interactions, and erosion of trust — if not carefully safeguarded. The ethical imperative that SDOH capture be accompanied by meaningful resource provision to address identified needs (the "bridge to nowhere" problem) applies with particular force in AI-mediated contexts.

6.4. Empirical Evidence on SDOH-Enhanced Models

Empirical evaluation of SDOH-enhanced AI triage remains in its infancy. Abbott et al. (2024) identified a predominance of proof-of-concept studies, with most focused on NLP extraction rather than on integration of SDOH variables into downstream predictive models. Available studies generally report modest but meaningful improvements in discrimination when SDOH features are added to clinical variables — typically in the range of 0.01–0.03 in AUROC — but the clinical significance of such gains is often unclear, and improvements in subgroup performance (which matter more for equity) are inconsistently reported. A further limitation is that most SDOH-augmented models have been developed and validated in single institutions, limiting generalizability.

6.5. The Paradox of SDOH Algorithmization

A critical concern that cuts across all dimensions of SDOH integration is what might be termed the paradox of algorithmization: tools intended to illuminate and address structural disadvantages may, if poorly designed, instead entrench them. Several mechanisms produce this risk. First, if SDOH variables are used as inputs to models that allocate scarce resources, patients with adverse social circumstances may systematically receive different — and potentially worse — treatment recommendations, reinforcing existing disparities. Second, treating SDOH as predictive features in clinical algorithms risks medicalizing structural inequality, shifting focus from upstream policy interventions to downstream risk stratification. Third, the labels on which SDOH-augmented models are trained often reflect disparities in the care received, not the care needed, creating the same proxy-variable problem identified by Obermeyer et al. (2019).

These concerns do not argue against SDOH integration in AI triage but do demand that it be pursued with methodological sophistication and ethical vigilance. Promising directions include causally informed modeling that distinguishes SDOH effects on need from SDOH effects on access; deployment pipelines that couple risk identification with automated resource linkage; and participatory design approaches that engage affected communities in the specification and evaluation of SDOH-enhanced algorithms. The goal, ultimately, is not to predict disadvantage but to identify actionable opportunities to mitigate it — a reframing that has profound implications for the selection of modeling targets, features, and evaluation metrics.

7. Ethical and Regulatory Frameworks and Bias Mitigation Strategies

7.1. Bioethical Principles in AI-Assisted Emergency Triage

The four classical principles of biomedical ethics — autonomy, beneficence, non-maleficence, and justice — acquire distinctive contours when applied to AI-assisted ED triage. *Autonomy* is complicated by the time-pressured, often acute nature of ED presentations in which informed consent to algorithmic assessment is practically impossible; transparency to patients about the role of AI in their care, and meaningful opportunity for clinician override, become proxies for autonomy in this context. *Beneficence* and *non-maleficence* require attention not only to average model performance but also to worst-case subgroup performance, since a model that improves overall accuracy while systematically harming a vulnerable minority violates the duty not to harm. *Justice* — the equitable distribution of benefits and burdens — is the principle most directly implicated by algorithmic bias and SDOH integration, requiring explicit attention to whether AI-enabled triage reduces or exacerbates health inequities. The WHO's 2021 guidance *Ethics and governance of artificial intelligence for health* articulated six consensus principles — protecting autonomy; promoting human well-being and safety; ensuring transparency; fostering responsibility and accountability; ensuring inclusiveness and equity; and promoting responsive and sustainable AI — that provide an overarching ethical scaffold for health AI (World Health Organization, 2021). The WHO's 2024 update, focused on large multi-modal models, reaffirms these principles while addressing the distinctive risks of generative AI in healthcare (World Health Organization, 2024).

7.2. Explainable AI: Promises and Limitations

Explainability — the capacity of a model to expose the reasoning behind its predictions to clinicians and patients — has become a central concern in health AI. Two techniques dominate current practice: *SHAP (SHapley Additive exPlanations)*, grounded in cooperative game theory, assigns each feature a contribution to an individual prediction, with the theoretical advantages of consistency and local accuracy; *LIME (Local*

Interpretable Model-agnostic Explanations) approximates complex models with simpler, locally faithful surrogates (Salih et al., 2025). Both methods have been widely applied in ED triage models, typically to identify which features most influence individual predictions of admission, ICU transfer, or mortality.

Despite their popularity, these post-hoc explanation methods carry significant limitations. Ghassemi et al. (2021), in an influential critique, argued that current explainable AI approaches offer a "false hope" in healthcare settings: explanations may be unstable across similar inputs, sensitive to feature collinearity, and vulnerable to adversarial manipulation. Yet SHAP and LIME explain what the model does but not whether what the model does is correct or clinically appropriate. An explanation that highlights ZIP code as a key predictor, for example, may reveal proxy bias rather than provide genuine clinical insight. A growing literature therefore emphasizes the complementary use of inherently interpretable models (e.g., generalized additive models, sparse rule lists) where appropriate, and the importance of causal reasoning in the validation of feature-importance attributions (Salih et al., 2025).

7.3. Regulatory Landscape

The regulatory environment for AI in clinical care has evolved substantially since 2024, driven by three major frameworks. The *EU AI Act* (Regulation 2024/1689), which entered into force in August 2024, establishes a risk-based classification system in which AI systems embedded in or constituting medical devices regulated under the Medical Device Regulation (2017/745) are automatically classified as high-risk (Aboy et al., 2024). Emergency patient triage systems are explicitly listed under Annex III as high-risk AI systems even outside the MDR context. High-risk systems must meet stringent requirements for risk management, data governance, technical documentation, transparency, human oversight, accuracy, robustness, and cybersecurity, with most obligations becoming applicable from August 2026 and full compliance required by August 2027.

The *US Food and Drug Administration* has pursued a more incremental regulatory path. Its 2021 *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device Action Plan* was followed by the *Good Machine Learning Practice for Medical Device Development: Guiding Principles*, jointly issued with the UK MHRA and Health Canada in October 2021 and updated in 2025. Subsequent guidance documents have addressed transparency, predetermined change control plans, and marketing submission expectations for AI-enabled device software functions (US Food and Drug Administration, 2024). The FDA's approach has emphasized total product lifecycle oversight, recognizing that AI systems may evolve after market authorization and require ongoing monitoring.

At the global level, the WHO's 2021 and 2024 guidance documents provide ethical and governance principles intended to inform national regulatory development, with particular attention to the needs of low- and middle-income countries where regulatory infrastructure may be less developed (World Health Organization, 2024). The Global Initiative on AI for Health (GI-AI4H), launched in 2023 by WHO, ITU, and WIPO, seeks to harmonize technical standards and policy guidance internationally.

7.4. Bias Mitigation Strategies

Technical and organizational strategies for bias mitigation span the AI lifecycle. *Pre-processing* interventions operate on training data, for example through reweighting to correct for underrepresentation of minority groups, synthesis of additional data for rare subgroups, or removal of biased proxy variables. *In-processing* interventions modify the learning procedure itself, for example through adversarial debiasing, fairness-constrained optimization, or multi-task objectives that jointly maximize accuracy and minimize disparity. *Post-processing* interventions adjust model outputs after training — for example, by applying group-specific decision thresholds or recalibration — to achieve chosen fairness criteria (Chen et al., 2023).

Algorithmic stewardship has emerged as a complementary organizational paradigm, emphasizing continuous post-deployment monitoring of model performance across demographic subgroups, prospective audits for drift and bias, incident reporting, and regular retraining with updated data. Federated learning offers a parallel technical frontier: by training shared models across institutions without pooling patient data, it enables broader representation of diverse populations while preserving privacy, and recent work has extended federated learning to incorporate fairness objectives explicitly (Zhang et al., 2025).

Participatory design — engaging patients, community members, and frontline clinicians in the specification, development, and evaluation of AI triage systems — represents a further strategy that has gained traction. Abbott et al. (2024) and others have argued that meaningful inclusion of affected communities is essential to identify biased proxies, clinically inappropriate features, and implementation pitfalls that purely technical audits may miss. Such participatory approaches reflect a broader shift from viewing fairness as a technical constraint to understanding it as a sociotechnical property emerging from the interaction of algorithm, workflow, and institution.

Finally, transparent reporting — exemplified by the TRIPOD+AI (Collins et al., 2024) and CONSORT-AI reporting guidelines — underpins all other mitigation strategies. Without complete, consistent reporting of model development, validation, and subgroup performance, external evaluation and scrutiny are impossible, and harmful systems may propagate undetected. Adherence to such standards remains imperfect but is slowly improving, driven by journal policies, regulatory expectations, and evolving community norms.

8. Discussion

8.1. Synthesis of Principal Findings

This narrative review has traced three interconnected strands of evidence regarding AI in emergency department triage. First, the technical literature consistently demonstrates that machine-learning models — particularly gradient-boosted ensembles and NLP-augmented architectures — can achieve substantially better discrimination of critical clinical endpoints than conventional five-level triage scales. Typical AUROC gains of 0.05 to 0.10 over ESI, MTS, or CTAS, combined with larger absolute gains in precision–recall performance for imbalanced outcomes, translate into meaningful improvements in the identification of critically ill patients (Almulihi et al., 2024; Grant et al., 2025; Sitthiprawiat et al., 2025). Recent prospective evaluations, notably that of Taylor et al. (2025), have begun to demonstrate that these technical gains can translate into real-world improvements in triage sensitivity and patient flow when AI is deployed as clinical decision support rather than as autonomous triage.

Second, however, the fairness literature documents substantial risks that AI triage systems may encode, propagate, or even amplify existing health inequities. These risks arise through multiple pathways — biased proxy variables, unrepresentative training data, measurement bias in physiological signals, and biased outcome labels — and are most dangerous when they remain unexamined. The Obermeyer et al. (2019) study remains the clearest demonstration that algorithms can produce systematic racial disparities even when race is excluded from model inputs.

Third, the integration of social determinants of health into AI triage holds both promise and peril. SDOH-augmented models may improve predictive accuracy and, more importantly, identify opportunities for targeted intervention; but without careful design, they risk medicalizing structural inequality and stigmatizing vulnerable patients (Abbott et al., 2024). The evidence base on operational SDOH integration in ED AI remains thin, with a predominance of proof-of-concept studies and limited external validation.

Emerging evidence suggests that carefully designed AI triage systems can reduce, rather than exacerbate, disparities. Hinson, Levin, et al. (2025) reported that a multisite AI-assisted triage implementation reduced undertriage disparities across racial and ethnic groups, offering preliminary evidence that AI can function as a corrective rather than reinforcing instrument when equity is an explicit design consideration.

8.2. Tensions, Trade-offs, and Unresolved Questions

Several fundamental tensions run through this field. The first concerns the trade-off between global accuracy and subgroup fairness. Optimizing for overall discrimination can produce models that perform poorly for minority subgroups, while enforcing fairness criteria can modestly reduce aggregate performance. The appropriate balance is not a technical question alone but a normative judgment about which form of error is more acceptable. The mathematical impossibility of simultaneously satisfying all common fairness criteria, demonstrated by Kleinberg et al. (2017), means that every deployed system implicitly embodies choices about whose interests are prioritized — choices that warrant explicit articulation and democratic legitimacy.

The second tension concerns explainability and performance. Deep learning and complex ensemble methods generally achieve the best discrimination but are the least interpretable; inherently interpretable models are often less accurate but more amenable to clinical scrutiny. Post-hoc explanation methods such as SHAP and LIME attempt to bridge this gap but have substantial limitations (Ghassemi et al., 2021; Salih et al., 2025). In high-stakes settings such as triage, the appropriate choice may depend on the nature of the decision, the reversibility of its consequences, and the capacity for meaningful human oversight.

The third tension is between data richness and privacy. More granular data — including social determinants, behavioral histories, and fine-grained physiological signals — may improve model performance but expand the attack surface for privacy breaches and stigmatization. Federated learning offers a partial technical solution, but organizational, legal, and ethical challenges to its adoption remain substantial (Zhang et al., 2025).

The fourth tension, perhaps most fundamental, concerns innovation and regulation. Excessive regulatory burden may stifle beneficial innovation, particularly at smaller institutions serving vulnerable populations; insufficient oversight may permit harmful systems to proliferate. The EU AI Act and related frameworks

represent the most ambitious attempts to navigate this trade-off, but their real-world impact on equity, innovation, and clinical utility will take years to assess.

8.3. Implications for Clinical Practice and Policy

Several implications follow from this synthesis. For clinicians, the central message is that AI triage tools should be understood as decision support, not decision replacement. The evidence that nurse override of AI-generated scores can improve validity in some settings, and that AI can complement nurse judgment in others (Taylor et al., 2025), underscores the importance of maintaining human authority at the triage bedside. Clinical adoption should be conditional on clear evidence of subgroup performance, transparent documentation of limitations, and pathways for reporting anomalies.

For health systems, the implications extend beyond individual tool adoption to include governance infrastructure: algorithmic impact assessments, inventories of deployed AI systems, continuous performance monitoring, and mechanisms for patient and community input. The concept of algorithmic stewardship — paralleling the more familiar antimicrobial stewardship — provides a useful organizational frame. Institutions should be prepared not only to deploy AI but also to retire or modify systems when monitoring reveals harm.

For policymakers and regulators, the priority agenda includes harmonizing international standards where possible, ensuring that regulatory frameworks attend to equity and not merely safety, and creating infrastructure to enable post-market surveillance of AI systems across diverse populations. The European Health Data Space initiative illustrates the potential for regulated data-sharing infrastructures to support both innovation and accountability (Aboy et al., 2024).

8.4. Limitations of This Review

Several limitations of this review warrant explicit acknowledgment. As a narrative synthesis, it does not involve formal risk-of-bias assessment of individual studies or quantitative meta-analysis, limiting the precision of comparative claims. The literature on AI in ED triage is evolving rapidly, and a narrative review inevitably represents a snapshot; novel architectures (including large language models), prospective trials, and regulatory developments may modify the conclusions drawn here. The review focuses predominantly on English-language literature and on studies conducted in high-income settings, limiting insight into low- and middle-income countries where ED crowding, inequity, and potential AI benefits may be even more pronounced. Finally, publication bias — the preferential publication of studies reporting favorable algorithmic performance — likely influences the evidence base, inflating the apparent gains of AI over conventional triage.

8.5. Directions for Future Research

The research agenda emerging from this synthesis is broad but focused. First, the field urgently needs prospective multicenter implementation trials that evaluate not only discrimination and calibration but also subgroup performance, clinician behavior, patient outcomes, and equity-related endpoints. Taylor et al. (2025) provide a model, but broader replication is needed across diverse settings. Second, systematic external validation and post-deployment monitoring should become routine rather than exceptional; registries of deployed AI triage systems, akin to device registries, would facilitate real-world learning. Third, federated learning offers a promising framework for developing models on diverse datasets while preserving privacy, and its combination with fairness-aware objectives merits sustained investigation (Zhang et al., 2025). Fourth, standardization of SDOH data capture — building on the work of the Gravity Project and the PRAPARE and AHC-HRSN instruments — is essential for meaningful integration of social determinants into triage models. Fifth, research targeting populations currently underrepresented in AI development, including rural, non-English-speaking, and LMIC patient groups, is a critical equity imperative. Sixth, the emerging use of large language models in ED triage, including through tools such as clinical chatbots and symptom checkers, warrants careful evaluation with respect to both performance and safety, informed by the TRIPOD-LLM reporting framework and WHO's 2024 guidance on large multi-modal models. Finally, interdisciplinary research that integrates ethnographic, qualitative, and participatory methods with quantitative model evaluation is essential to illuminate the sociotechnical dynamics that determine whether AI triage improves or worsens care in practice.

9. Conclusions

Artificial intelligence now plays a growing role in emergency department triage. The technical evidence is increasingly robust: AI models can identify critically ill patients more accurately than conventional five-level triage scales, predict hospital and ICU admission with high discrimination, and capture clinical nuance from unstructured text that structured fields miss. Early prospective evaluations suggest that these technical gains can translate into measurable improvements in real-world triage performance and, under the right conditions, into reductions in the demographic disparities that have long characterized ED triage.

At the same time, the fairness and equity literature documents clear risks. Algorithmic triage systems can encode and amplify biases present in their training data, measurement instruments, and outcome labels. The Obermeyer et al. (2019) study of commercial care-management algorithms, combined with subsequent work on pulse oximetry bias, race-based calculators, and the limits of standard fairness metrics, establishes that equity cannot be assumed as a byproduct of accuracy; it must be designed, monitored, and enforced. The integration of social determinants of health introduces further opportunities and risks, promising more holistic risk stratification but raising substantive concerns about stigmatization, medicalization of structural inequality, and the adequacy of current data infrastructures.

Three broad recommendations follow for the main stakeholder communities. For researchers, the priority is rigorous, transparent evidence generation: prospective trials with subgroup analyses, consistent adherence to TRIPOD+AI and related reporting standards, investment in external validation, and meaningful engagement with affected communities in model design. For clinicians and health systems, the priority is cautious but purposeful adoption: deploying AI as decision support, instituting algorithmic stewardship, monitoring for disparate performance, and preserving human authority in high-stakes decisions. For policymakers and regulators, the priority is development of governance frameworks — exemplified by the EU AI Act and WHO's evolving guidance — that safeguard equity alongside safety, support post-market surveillance, and facilitate responsible international harmonization.

The ultimate test of AI in ED triage will not be its peak discriminative accuracy in a benchmark dataset but its sustained ability to improve the quality and equity of care delivered to the most diverse patients at their moments of greatest vulnerability. Meeting this test requires combining technical innovation with ethical vigilance, regulatory maturity, and a sustained commitment to health equity. The challenge is substantial, but so is the opportunity: against rising demand for emergency care, persistent health inequities, and rapid technological change, careful development of AI-assisted triage offers a major opportunity to advance both safety and fairness in emergency medicine.

Author Contributions

Conceptualization and methodology: Damian Danilczuk.
Validation: Wiktor Adamiec, Przemysław Piterak, Adrianna Buż.
Formal analysis: Damian Danilczuk, Michał Dyś, Monika Szlachta-Gubernat.
Resources: Jerzy Buszko, Marta Bajkowska-Piterak.
Writing – original draft preparation: Damian Danilczuk, Nicol Baran.
Writing – review and editing: Marta Bajkowska-Piterak, Monika Szlachta-Gubernat.
Supervision: Jerzy Buszko.
Project administration: Damian Danilczuk, Przemysław Piterak.

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

Conflicts of Interest: The authors declare no conflict of interest.

Funding: This research received no external funding.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process: No artificial intelligence tools were used to generate the scholarly content of this review; AI-assisted tools were limited to language editing and formatting support, with all substantive analysis, interpretation, and argumentation performed by the authors.

REFERENCES

1. Abbott, E. E., Apakama, D., Richardson, L. D., Chan, L., & Nadkarni, G. N. (2024). Leveraging artificial intelligence and data science for integration of social determinants of health in emergency medicine: Scoping review. *JMIR Medical Informatics, 12*, e57124. https://doi.org/10.2196/57124
2. Aboy, M., Minssen, T., & Vayena, E. (2024). Navigating the EU AI Act: Implications for regulated digital medical products. *npj Digital Medicine, 7*, 237. https://doi.org/10.1038/s41746-024-01232-3
3. Almulihi, Q. A., Alquraini, A. A., Almulihi, F. A. A., Alzahid, A. A., Al Qahtani, S. S. A. J., Almulhim, M., Alqhtani, S. H. S., Alnafea, F. M. N., Mushni, S. A. S., Alaqil, N. A., Assiri, M. I. F., & Maghraby, N. H. (2024). Applications of artificial intelligence and machine learning in emergency medicine triage: A systematic review. *Medical Archives, 78*(3), 198–206. https://doi.org/10.5455/medarh.2024.78.198-206
4. Baethge, C., Goldbeck-Wood, S., & Mertens, S. (2019). SANRA—A scale for the quality assessment of narrative review articles. *Research Integrity and Peer Review, 4*, 5. https://doi.org/10.1186/s41073-019-0064-8
5. Bullard, M. J., Musgrave, E., Warren, D., Unger, B., Skeldon, T., Grierson, R., van der Linde, E., & Swain, J. (2017). Revisions to the Canadian Emergency Department Triage and Acuity Scale (CTAS) guidelines 2016. *CJEM, 19*(S2), S18–S27. https://doi.org/10.1017/cem.2017.365
6. Cairns, C., Ashman, J. J., & King, J. M. (2023). *Emergency department visit rates by selected characteristics: United States, 2021* (NCHS Data Brief No. 478). National Center for Health Statistics.
7. Chang, C. Y., Abujaber, S., Reynolds, T. A., Camargo, C. A., Jr., & Obermeyer, Z. (2016). Burden of emergency conditions and emergency care usage: New estimates from 40 countries. *Emergency Medicine Journal, 33*(11), 794–800. https://doi.org/10.1136/emermed-2016-205709
8. Chang, H., Yu, J. Y., Yoon, S., Kim, T., & Cha, W. C. (2022). Machine learning-based suggestion for critical interventions in the management of potentially severe conditioned patients in emergency department triage. *Scientific Reports, 12*, 10537. https://doi.org/10.1038/s41598-022-14422-4
9. Chen, P., Wu, L., & Wang, L. (2023). AI fairness in data management and analytics: A review on challenges, methodologies and applications. *Applied Sciences, 13*(18), 10258. https://doi.org/10.3390/app131810258
10. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. https://doi.org/10.1136/bmj-2023-078378
11. El Arab, R. A., & Al Moosa, O. A. (2025). The role of AI in emergency department triage: An integrative systematic review. *Intensive and Critical Care Nursing, 89*, 104058. https://doi.org/10.1016/j.iccn.2025.104058
12. Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing, 24*(4), 230–235. https://doi.org/10.1179/2047480615Z.000000000329
13. Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9
14. Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine, 178*(11), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763
15. Gräff, I., Goldschmidt, B., Glien, P., Bogdanow, M., Fimmers, R., Hoeft, A., Kim, S.-C., & Grigutsch, D. (2014). The German version of the Manchester Triage System and its quality criteria—First assessment of validity and reliability. *PLOS ONE, 9*(2), e88995. https://doi.org/10.1371/journal.pone.0088995
16. Grant, L., Diagne, M., Aroutiunian, R., Hopkins, D., Bai, T., Kondrup, F., & Clark, G. (2025). Machine learning outperforms the Canadian Triage and Acuity Scale (CTAS) in predicting need for early critical care. *Canadian Journal of Emergency Medicine, 27*(1), 43–52. https://doi.org/10.1007/s43678-024-00807-z
17. Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal, 26*(2), 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x
18. Greenhalgh, T., Thorne, S., & Malterud, K. (2018). Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation, 48*(6), e12931. https://doi.org/10.1111/eci.12931
19. Hall, J. N., McCarron, J., Toarta, C., McLeod, S. L., on behalf of the CTAS National Advisory Committee and the NENA Triage Committee. (2025). Canadian Emergency Department Triage and Acuity Scale (CTAS) guidelines 2025. *Canadian Journal of Emergency Medicine, 27*, 774–777. https://doi.org/10.1007/s43678-025-00996-1
20. 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

21. Hinson, J. S., Martinez, D. A., Schmitz, P. S. K., Toerper, M., Radu, D., Scheulen, J., Stewart de Ramirez, S. A., & 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. *International Journal of Emergency Medicine, 11*(1), 3. https://doi.org/10.1186/s12245-017-0161-8
22. Hong, W. S., Haimovich, A. D., & Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning. *PLOS ONE, 13*(7), e0201016. https://doi.org/10.1371/journal.pone.0201016
23. Hood, C. M., Gennuso, K. P., Swain, G. R., & Catlin, B. B. (2016). County health rankings: Relationships between determinant factors and health outcomes. *American Journal of Preventive Medicine, 50*(2), 129–135. https://doi.org/10.1016/j.amepre.2015.08.024
24. Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou (Ed.), *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)* (Vol. 67, pp. 43:1–43:23). Schloss Dagstuhl–Leibniz-Zentrum für Informatik. https://doi.org/10.4230/LIPIcs.ITCS.2017.43
25. Mirhaghi, A., Heydari, A., Mazlom, R., & Ebrahimi, M. (2015). The reliability of the Canadian Triage and Acuity Scale: Meta-analysis. *North American Journal of Medical Sciences, 7*(7), 299–305. https://doi.org/10.4103/1947-2714.161243
26. Mirhaghi, A., Mazlom, R., Heydari, A., & Ebrahimi, M. (2017). The reliability of the Manchester Triage System (MTS): A meta-analysis. *Journal of Evidence-Based Medicine, 10*(2), 129–135. https://doi.org/10.1111/jebm.12231
27. Moen, M., Storr, C., German, D., Friedmann, E., & Johantgen, M. (2020). A review of tools to screen for social determinants of health in the United States: A practice brief. *Population Health Management, 23*(6), 422–429. https://doi.org/10.1089/pop.2019.0158
28. Morley, C., Unwin, M., Peterson, G. M., Stankovich, J., & Kinsman, L. (2018). Emergency department crowding: A systematic review of causes, consequences and solutions. *PLOS ONE, 13*(8), e0203316. https://doi.org/10.1371/journal.pone.0203316
29. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342
30. Pearce, S., Marchand, T., Shannon, T., Ganshorn, H., & Lang, E. (2023). Emergency department crowding: An overview of reviews describing measures, causes, and harms. *Internal and Emergency Medicine, 18*(4), 1137–1158. https://doi.org/10.1007/s11739-023-03239-2
31. Porto, B. M. (2024). Improving triage performance in emergency departments using machine learning and natural language processing: A systematic review. *BMC Emergency Medicine, 24*(1), 219. https://doi.org/10.1186/s12873-024-01135-2
32. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of Internal Medicine, 169*(12), 866–872. https://doi.org/10.7326/M18-1990
33. Rasouli, H. R., Aliakbar Esfahani, A., Nobakht, M., Eskandari, M., Mahmoodi, S., Goodarzi, H., & Abbasi Farajzadeh, M. (2019). Outcomes of crowding in emergency departments: A systematic review. *Archives of Academic Emergency Medicine, 7*(1), e52.
34. Salih, A. M., Raisi-Estabragh, Z., Galazzo, I. B., Radeva, P., Petersen, S. E., Lekadir, K., & Menegaz, G. (2025). A perspective on explainable artificial intelligence methods: SHAP and LIME. *Advanced Intelligent Systems, 7*(1), 2400304. https://doi.org/10.1002/aisy.202400304
35. 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
36. Sitthiprawiat, P., Wittayachamnankul, B., Sirikul, W., & Laohavisudhi, K. (2025). Development and internal validation of an AI-based emergency triage model for predicting critical outcomes in emergency department. *Scientific Reports, 15*, 31212. https://doi.org/10.1038/s41598-025-17180-1
37. Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial bias in pulse oximetry measurement. *New England Journal of Medicine, 383*(25), 2477–2478. https://doi.org/10.1056/NEJMc2029240
38. Sukhera, J. (2022). Narrative reviews: Flexible, rigorous, and practical. *Journal of Graduate Medical Education, 14*(4), 414–417. https://doi.org/10.4300/JGME-D-22-00480.1
39. 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
40. Truong, H. P., Luke, A. A., Hammond, G., Wadhera, R. K., Reidhead, M., & Joynt Maddox, K. E. (2020). Utilization of social determinants of health ICD-10 Z-codes among hospitalized patients in the United States, 2016–2017. *Medical Care, 58*(12), 1037–1043. https://doi.org/10.1097/MLR.0000000000001418
41. Tyler, S., Olis, M., Aust, N., Patel, L., Simon, L., Triantafyllidis, C., Patel, V., Lee, D. W., Ginsberg, B., Ahmad, H., & Jacobs, R. J. (2024). Use of artificial intelligence in triage in hospital emergency departments: A scoping review. *Cureus, 16*(5), e59906. https://doi.org/10.7759/cureus.59906

42. U.S. Food and Drug Administration. (2024). *Marketing submission recommendations for a predetermined change control plan for artificial intelligence-enabled device software functions: Guidance for industry and Food and Drug Administration staff*. U.S. Department of Health and Human Services. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/marketing-submission-recommendations-predetermined-change-control-plan-artificial-intelligence
43. Vyas, D. A., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight—Reconsidering the use of race correction in clinical algorithms. *New England Journal of Medicine, 383*(9), 874–882. https://doi.org/10.1056/NEJMms2004740
44. World Health Organization. (2010). *A conceptual framework for action on the social determinants of health* (Social Determinants of Health Discussion Paper 2). WHO.
45. World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. WHO.
46. World Health Organization. (2024). *Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models*. WHO.
47. Yun, H., Choi, J., & Park, J. H. (2021). Prediction of critical care outcome for adult patients presenting to emergency department using initial triage information: An XGBoost algorithm analysis. *JMIR Medical Informatics, 9*(9), e30770. https://doi.org/10.2196/30770
48. Zachariasse, J. M., van der Hagen, V., Seiger, N., Mackway-Jones, K., van Veen, M., & Moll, H. A. (2019). Performance of triage systems in emergency care: A systematic review and meta-analysis. *BMJ Open, 9*(5), e026471. https://doi.org/10.1136/bmjopen-2018-026471
49. Zhang, F., Zhai, D., Bai, G., Jiang, J., Ye, Q., Ji, X., & Liu, X. (2025). Towards fairness-aware and privacy-preserving enhanced collaborative learning for healthcare. *Nature Communications, 16*, 2852. https://doi.org/10.1038/s41467-025-58055-3