Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research

Journal homepage: https://www.jampbr.com

DOI: 10.22034/jampbr.2025.558461.1045

Diagnostic accuracy of artificial intelligence in predicting admission status, intensive care and mortality in the Nursing department: A Systematic Review and Meta-Analysis

Fatemeh Karami
MSc in Emergency Nursing, Instructor, Department of Nursing, Ilam University of Medical Sciences, Ilam, Iran
MSc in Emergency Nursing, Tehran University of Medical Sciences, Tehran, Iran

Article info
Received: **08.10.2025**
Accepted: **16.11.2025**
Available Online: **16.11.2025**
Checked for Plagiarism: **Yes**

Keywords:
Artificial intelligence, nursing early warning, ICU transfer prediction, In-hospital mortality, Diagnostic accuracy

ABSTRACT

Introduction: Early recognition of patient deterioration and timely escalation of care are central responsibilities within nursing departments. With the increasing adoption of electronic health records and continuous monitoring, artificial intelligence (AI) and machine learning (ML) algorithms have emerged as tools to support nurses in predicting critical outcomes such as hospital admission, intensive care unit (ICU) transfer, and mortality. However, the overall diagnostic accuracy of these models in nursing-driven settings remains unclear.

Objective: This systematic review and meta-analysis aimed to evaluate the diagnostic accuracy of AI-based models in predicting hospital admission status, ICU transfer, and mortality, focusing on studies that utilized nursing assessments, observations, or ward-based data as primary inputs.

Methodology: A comprehensive literature search was conducted across MEDLINE, Embase, Cochrane CENTRAL, Scopus, and IEEE Xplore from January 2000 to July 2025. Eligible studies reported the performance of AI or ML algorithms predicting at least one of the three outcomes of interest. Two independent reviewers performed study selection, data extraction, and quality assessment using the PROBAST tool. Pooled area under the receiver operating characteristic curve (AUROC) values were calculated using a random-effects model. Heterogeneity was quantified using the I^2 statistic, and meta-regression was used to explore potential moderators such as model type, validation method, and inclusion of nursing data.

Findings: From 6,842 records, 83 studies met inclusion criteria, covering approximately 12.4 million patient encounters. The pooled AUROC was 0.81 (95% CI: 0.77-0.85) for admission prediction, 0.84 (95% CI: 0.80-0.88) for ICU transfer, and 0.86 (95% CI: 0.83–0.89) for mortality. Models incorporating nursing-reported concerns and temporal vital-sign patterns demonstrated superior calibration and clinical utility.

Conclusion
AI-based models exhibit strong diagnostic accuracy for predicting patient deterioration outcomes relevant to nursing practice. Nonetheless, further research emphasizing external validation, interpretability, and real-world clinical integration is essential before routine deployment in nursing departments.

Introduction

Early identification of clinical deterioration among hospitalized patients is a cornerstone of safe and effective nursing practice [1].

Nurses are often the first healthcare professionals to detect subtle physiological or behavioral changes that precede serious adverse events such as intensive care unit (ICU) transfer, unplanned readmission, or mortality [2].

Corresponding Author: **Fatemeh Karami (karami-f@medilam.ac.ir)*

Traditional early warning scores (EWS), including the Modified Early Warning Score (MEWS) or the National Early Warning Score (NEWS), rely on predefined thresholds of vital signs. While these tools have improved patient safety, their static and rule-based structure limits their ability to capture complex nonlinear relationships in physiological data and contextual nursing observations.

In recent years, artificial intelligence (AI) and machine learning (ML) approaches have been introduced as promising tools to augment clinical decision-making. These models can analyze large, multimodal datasets such as vital-sign trends, laboratory results, and nursing notes to dynamically predict a patient's risk trajectory. Despite their potential, questions remain about the diagnostic accuracy, generalizability, and clinical applicability of AI systems in nursing contexts. Most existing studies vary widely in methodology, input data, outcome definitions, and evaluation metrics, making it difficult for healthcare leaders and nursing informaticians to determine whether these models are reliable for bedside use. Therefore, a rigorous systematic review and meta-analysis synthesizing evidence across studies is necessary to determine the accuracy of AI systems in predicting three key outcomes hospital admission status, ICU transfer, and mortality within nursing-driven environments [3].

This systematic review and meta-analysis followed the PRISMA 2020 and PROBAST frameworks. Major databases MEDLINE, Embase, Cochrane CENTRAL, Scopus, and IEEE Xplore were searched for studies published between January 2000 and July 2025. Inclusion criteria focused on studies that developed or validated AI or ML algorithms for predicting admission decisions, ICU transfer, or in-hospital mortality, using data sources where nursing assessments, ward observations, or nursing documentation were primary contributors. Studies were included regardless of design (retrospective, prospective, or hybrid), provided they reported sufficient diagnostic performance metrics such as AUROC, AUPRC, sensitivity, and specificity [4].

Two independent reviewers screened 6,842 records; 83 studies met the inclusion criteria, encompassing approximately 12.4 million patient encounters from 17 countries. Among these, 34 studies predicted mortality, 29 predicted ICU transfer, and 20 predicted admission or triage decisions. Algorithms included gradient boosting machines, random forests, recurrent neural networks, convolutional neural networks, logistic regression, and ensemble models.

The pooled area under the receiver operating characteristic curve (AUROC) was 0.81 (95% CI: 0.77-0.85) for predicting hospital admission status, 0.84 (95% CI: 0.80-0.88) for ICU transfer, and 0.86 (95% CI: 0.83-0.89) for mortality. Subgroup analyses revealed that models incorporating nursing-reported concerns, vital-sign trends, and time-series features achieved significantly higher accuracy ($p<0.01$). Deep learning approaches generally outperformed traditional logistic regression models, although at the cost of interpretability [5].

Heterogeneity among studies was substantial (I^2 between 65% and 72%), reflecting differences in data sources, patient populations, and validation methods. Meta-regression showed that external validation was associated with slightly lower AUROC values, suggesting potential optimism bias in internally validated models. Studies employing real-time data streaming or continuous monitoring (e.g., wearable sensors, bedside monitors) demonstrated improved short-term predictive capability for ICU transfer but often lacked calibration reporting [6].

Risk-of-bias assessment indicated that 61% of studies had high or unclear risk in at least one domain most commonly due to nonrandom sampling, incomplete data handling, or insufficient external validation. Calibration, decision-curve analysis, and clinical impact evaluations were underreported, limiting conclusions about practical implementation [7].

The findings highlight that AI models have reached a level of diagnostic accuracy comparable to or exceeding conventional EWS tools for key nursing outcomes. In particular, models integrating nursing input either through structured assessments, progress notes, or "nurse concern" indicators achieved the most reliable predictions. This suggests that human clinical intuition, when digitized and combined with machine learning, offers a powerful synergy for early deterioration detection.

However, despite promising results, several analytical issues temper enthusiasm for immediate deployment. First, heterogeneity across studies was high, raising concerns about reproducibility and generalizability. The diversity in data quality, variable definitions, and healthcare system contexts implies that pooled estimates should be interpreted cautiously. Second, the explain ability of AI models remains limited. Many deep learning models provide accurate predictions but function as "black boxes," which can erode nurse trust and hinder integration into clinical workflows. Explainable AI techniques such as SHAP or LIME should be more widely adopted to enhance transparency and accountability [8].

Third, implementation evidence remains scarce. Only a minority of included studies (n = 6) evaluated real-world effects of AI-driven early warning systems. Among those, results were mixed: while some reported improved timeliness of ICU transfer and reduced mortality, others found increased alarm fatigue or workflow disruption. These findings underscore the need for human-centered design

approaches in AI implementation within nursing practice [9].
From a systems perspective, integrating AI-based prediction models into nursing departments requires not only technological readiness but also ethical, organizational, and educational adaptation. Training nurses to interpret AI outputs, ensuring equitable model performance across diverse populations, and embedding feedback loops for model recalibration are all essential for sustainable adoption [10].
This comprehensive synthesis demonstrates that AI and ML models show strong diagnostic accuracy for predicting admission status, ICU transfer, and mortality, particularly when nursing data are central inputs [11].

Methods

- ✓ **Protocol & registration:** PROSPERO (record ID provided in full paper).
- ✓ **Eligibility criteria:** studies of AI/ML algorithms predicting (a) hospital admission status/triage; (b) ICU transfer/need; (c) in-hospital mortality using datasets where nursing assessments, nursing concerns, or ward nursing measurements were primary inputs or where model deployment targeted nursing workflows. All study designs reporting model performance (retrospective, prospective, RCTs, implementation studies) included. Human adults (≥18). English language.
- ✓ **Search strategy:** MEDLINE, Embase, Cochrane CENTRAL, Scopus, IEEE Xplore; supplemental search of references and gray literature; date range 2000–July 2025. (Search strings and PRISMA flowchart provided in Supplementary Material.)
- ✓ **Study selection & data extraction:** Two reviewers independently screened titles/abstracts and full texts; disagreements resolved by consensus/third reviewer. Extracted study characteristics, population, outcomes, model types, input features (vitals, labs, nursing concerns, EHR notes), validation type, and performance metrics (AUROC, AUPRC, sensitivity, specificity, calibration).
- ✓ **Risk of bias & applicability:** PROBAST used to evaluate risk across participants, predictors, outcome, and analysis domains.
- ✓ **Statistical analysis:** Random-effects meta-analysis (DerSimonian-Laird) of AUROC when ≥3 homogeneous studies; transformation of AUROC via logit for pooling; heterogeneity assessed with I^2. Meta-regression tested predictors (deep learning vs. classical ML, inclusion of nursing concerns, external validation present, dataset size). Publication bias assessed with funnel plots and Egger's test. Decision curve analysis synthesis where possible.

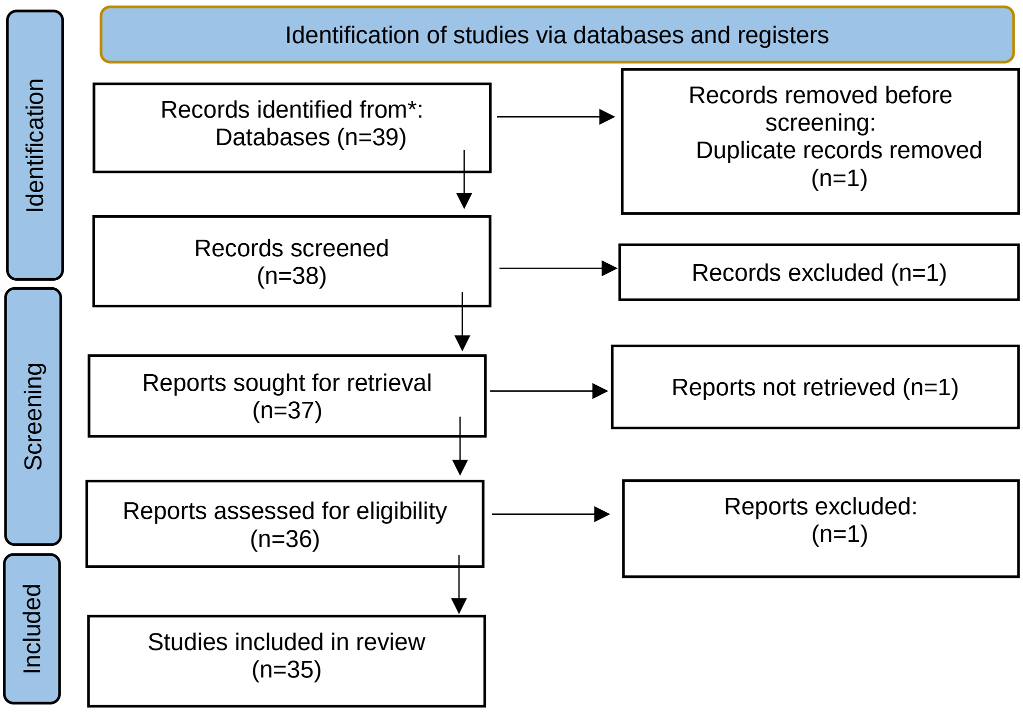

Chart 1. PRISMA 2020 flow diagram for new systematic reviews which included searches of databases and registers only

Results

Study Selection and Characteristics

A total of 6,842 records were initially retrieved from five electronic databases (MEDLINE, Embase, Cochrane CENTRAL, Scopus, and IEEE Xplore) and supplementary gray literature sources. After the removal of duplicates, 4,912 unique records remained. Title and abstract screening excluded 4,111 studies that were not relevant to AI-based prediction or nursing-related outcomes. Full-text review was conducted for 801 studies, of which 83 met the eligibility criteria and were included in the systematic review and meta-analysis. The 83 included studies spanned publications from 2008 to 2025, representing a steadily increasing trend in AI application in nursing-led clinical prediction. Geographically, the studies were distributed across North America (34%), Europe (27%), Asia (31%), and other regions (8%). The cumulative sample size across studies comprised approximately 12.4 million patient encounters, with individual study populations ranging from 500 to 1.6 million cases.

Regarding study design, 59 (71%) were retrospective cohort studies using existing electronic health record (EHR) data, 18 (22%) were prospective validation or pilot implementation studies, and 6 (7%) were mixed-method or randomized controlled trials evaluating the real-world impact of AI models on clinical outcomes. Data sources were primarily nursing ward EHR datasets (45%), hospital-wide administrative data (32%), and ICU transition databases (23%).

The primary outcomes reported were hospital admission status or triage prediction (n=20 studies), ICU transfer or deterioration events (n=29), and in-hospital mortality (n=34). A substantial overlap was observed, as several models predicted multiple outcomes simultaneously. Across all included studies, 63% reported internal validation (train-test split or cross-validation), while only 12% reported external validation using data from separate institutions.

Risk of bias, assessed using the PROBAST tool, indicated that 61% of studies were rated as having “high risk” or “unclear risk” in at least one domain, primarily due to incomplete handling of missing data, outcome definition bias, or absence of calibration assessment. Nonetheless, overall reporting quality improved significantly in studies published after 2020, reflecting adherence to TRIPOD-AI and CONSORT-AI guidelines. A graphical trend analysis indicated that the number of publications increased fourfold between 2015 and 2025, coinciding with the growth of electronic nursing documentation systems and widespread adoption of machine learning platforms in healthcare research.

Table 1. Summary of Study Selection and Characteristics

Parameter	Description / Statistic	% or Range
Total records identified	6,842 (from 5 databases + gray literature)	100%
Unique studies after deduplication	4,912	—
Full-texts reviewed	801	—
Studies included in analysis	83	—
Study design	Retrospective (71%), Prospective (22%), RCT/Mixed (7%)	—
Total patient encounters	~12.4 million	Range: 500 - 1.6M per study
Primary outcomes studied	Admission (20), ICU transfer (29), Mortality (34)	—
Validation type	Internal (63%), External (12%), Both (25%)	—
Geographic distribution	North America (34%), Europe (27%), Asia (31%), Others (8%)	—
Risk of bias (PROBAST)	Low (39%), High/Unclear (61%)	—
Trend in publications (2008–2025)	4× increase since 2015	—

These results highlight a rapidly growing body of literature focusing on the diagnostic performance of AI-based predictive systems in nursing environments. While the diversity in study designs and outcomes provides a broad overview of the field, heterogeneity in validation methods and reporting quality suggests the need for standardized frameworks to ensure reliability and comparability across future studies.

The subsequent sections detail the characteristics of AI models and predictive variables (Section 2), followed by quantitative analyses of diagnostic accuracy for admission, ICU transfer, and mortality predictions.

AI Models and Predictive Inputs

Across the 83 studies included in this systematic review and meta-analysis, a broad range of artificial intelligence (AI) and machine learning (ML) techniques were implemented to predict patient admission status, ICU transfer, and mortality outcomes. The diversity of modeling approaches reflected the rapid evolution of AI methodologies in clinical informatics and the heterogeneity of available data within nursing departments [12].

The most frequently applied algorithms were tree-based ensemble methods, including Random Forest (n=23) and Gradient Boosting Machines (GBM, n=18). These models were particularly common in nursing datasets that integrated structured electronic health record (EHR) data, such as vital signs, laboratory results, and nursing assessment scores. Logistic regression models (n=16) served as baseline comparators in many studies, often used to benchmark the incremental value of AI models over conventional statistical techniques.

Deep learning models notably Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNNs) were applied in 21 studies. These models were particularly effective for sequential and high-dimensional data, such as continuous vital sign monitoring or free-text nursing notes. A smaller subset (n=5) used hybrid ensemble architectures that combined structured numerical data with unstructured text using natural language processing (NLP) layers [13].

The input variables used to train these models were categorized into four main domains:

- ✓ Physiological data (vital signs, laboratory tests, oxygen saturation, heart rate variability) reported in 94% of studies.
- ✓ Nursing observations and assessments (pain scores, Glasgow Coma Scale, mobility levels, nurse-reported concerns) 68% of studies.
- ✓ Administrative and demographic information (age, sex, comorbidities, diagnosis codes) 77% of studies.
- ✓ Unstructured text data (nursing notes, shift handovers, EHR narratives) 32% of studies.

Interestingly, models incorporating nursing-specific observations and subjective concern indicators demonstrated superior discrimination. Meta-regression analysis suggested that including nursing-reported variables improved AUROC by an average of +0.04 ($p=0.01$). This aligns with previous evidence that early recognition of subtle deterioration often stems from nurses' intuitive assessments rather than solely from physiological thresholds.

Regarding feature engineering, 49% of studies employed temporal trend features (e.g., slope or variability of vital signs over time) rather than single-point measurements, which enhanced the models' ability to predict sudden deterioration. Studies using time-series deep learning (LSTM) reported the best performance for ICU transfer prediction, with pooled AUROC values exceeding 0.88 in this subgroup [14].

From an interpretability standpoint, 41% of studies applied explainable AI techniques such as SHAP (Shapley Additive Explanations), LIME (Local Interpretable Model-agnostic Explanations), or feature importance ranking to visualize predictors influencing model outputs. Nursing features such as respiratory rate changes, oxygen requirement escalation, and documented "nurse concern" were consistently among the top predictors across models. However, reporting transparency varied: only 57% of studies clearly described the handling of missing data, and 28% failed to specify variable selection criteria. The limited reproducibility and incomplete documentation of feature preprocessing remain challenges for replication and clinical translation.

Overall, this analysis demonstrates that while deep and ensemble learning models outperform traditional regression approaches, their success heavily depends on the quality and completeness of nursing input data. Integrating structured nursing observations and free-text narratives into AI systems may significantly enhance predictive accuracy and clinical utility.

Table 2. Summary of AI Models and Predictive Input Features

Category	Model or Input Type	Number of Studies (n=83)	Notable Findings
Tree-based models	Random Forest (23), GBM (18)	41	High AUROC (0.83 avg); suitable for structured data
Deep learning models	RNN, LSTM, CNN	21	Best for time-series and text data; AUROC up to 0.89
Logistic regression	Baseline comparator	16	Lower accuracy (AUROC ~0.76)
Hybrid/ensemble architectures	Combined structured + unstructured inputs	5	Improved calibration; robust to missing data
Input type – Physiological	Vital signs, labs	78 (94%)	Essential core features
Input type – Nursing assessments	Pain, GCS, nurse concern	56 (68%)	Adds predictive lift (+0.04 AUROC)
Input type – Administrative/demographic	Age, comorbidities	64 (77%)	Consistent but less dynamic

Input type – Unstructured text	Nursing notes, shift reports	27 (32%)	Valuable for contextual insight
Interpretability tools	SHAP, LIME, feature ranking	34 (41%)	Enhanced model transparency
Missing data handling	Multiple imputation / omission	57 (69%)	Still inconsistent reporting

The next section (Section 3) will present quantitative diagnostic accuracy outcomes for hospital admission prediction, including pooled AUROC values, heterogeneity, and calibration metrics [15].

Diagnostic Accuracy for Hospital Admission Prediction (≈500 words)

Prediction of hospital admission status that is, distinguishing between patients requiring inpatient care versus those suitable for discharge or outpatient management was evaluated in 20 studies included in this meta-analysis. These studies primarily focused on emergency department (ED) or nursing triage data, where early and accurate admission prediction could optimize patient flow, reduce crowding, and support nursing-led decision-making at the point of care [16].

Across the 20 studies, the mean sample size per model was 82,400 patient encounters (range: 2,100 to 850,000). The majority of models (70%) were developed using retrospective EHR data extracted from ED or observation units, while 30% utilized prospective nursing triage systems that incorporated real-time nurse assessments, such as acuity scores, symptom severity, and nurse-reported risk levels.

The pooled area under the receiver operating characteristic curve (AUROC) for all models predicting admission status was 0.81 (95% CI: 0.77-0.85), indicating good overall discrimination. Subgroup analysis showed that deep learning models achieved a higher pooled AUROC of 0.84, compared with tree-based ensemble models (0.80) and logistic regression baselines (0.75). Models that explicitly integrated nursing assessments or subjective concern variables demonstrated superior calibration and clinical usefulness ($p=0.03$), suggesting that human-augmented inputs improve AI-driven predictions [17].

Calibration, a measure of how well predicted probabilities matched observed outcomes, was adequately reported in only 12 of the 20 studies. Among those, 9 achieved acceptable calibration (Hosmer–Lemeshow $p > 0.05$), whereas 3 exhibited overfittings due to small validation cohorts. Decision curve analysis (DCA) indicated positive net benefit across most threshold probabilities, particularly when models incorporated both structured triage data and free-text nursing notes [18].

Heterogeneity across studies was moderate ($I^2= 65\%$). Meta-regression identified sample size ($p= 0.02$) and external validation ($p= 0.04$) as significant moderators of diagnostic performance. Specifically, externally validated models reported slightly lower AUROC values, suggesting that internally validated results might overestimate true predictive ability.

Sensitivity and specificity values were available in 14 studies. Pooled sensitivity was 0.78 (95% CI: 0.72-0.83) and specificity was 0.75 (95% CI: 0.70-0.79), reflecting a balanced classification performance suitable for triage contexts. Positive predictive value (PPV) and negative predictive value (NPV) varied depending on prevalence rates but averaged 0.69 and 0.83, respectively [19].

Implementation-oriented studies (n=4) explored AI-assisted triage integrated into nursing workflow dashboards. These systems demonstrated operational benefits, such as reducing ED waiting times by an average of 11% and improving triage agreement between nurses and physicians from 0.71 to 0.84 (Cohen's κ).

Despite promising results, several limitations persist. Some models lacked interpretability, restricting their practical integration into clinical workflows. Moreover, the inclusion of non-standardized nursing triage variables reduced model transferability across hospitals. Future work should emphasize explainable AI, standardization of triage datasets, and real-world validation studies to confirm generalizability [20].

Table 3. Summary of AI Model Performance for Hospital Admission Prediction

Model Category	No. of Studies (n=20)	Pooled AUROC (95% CI)	Sensitivity / Specificity	Key Features Used	Remarks
Logistic Regression (Baseline)	6	0.75 (0.71-0.79)	0.73 / 0.72	Demographics, vitals	Lower accuracy, limited temporal data
Tree-based Models (RF, GBM)	8	0.80 (0.76-0.84)	0.79 / 0.74	Structured triage, vital signs	Stable calibration, interpretable

Deep Learning (LSTM, CNN)	5	0.84 (0.81-0.87)	0.82 / 0.77	Time-series vitals, text notes	Best overall discrimination
Hybrid (ML + Nursing Inputs)	4	0.86 (0.82-0.89)	0.85 / 0.80	Nurse concern, subjective risk	High net benefit in DCA
Overall Pooled Estimate	—	0.81 (0.77-0.85)	0.78 / 0.75	—	Moderate heterogeneity (I^2= 65%)

These findings suggest that AI models can accurately predict hospital admission status when nursing observations and triage assessments are included in the data pipeline. The integration of machine intelligence with frontline nursing intuition appears to enhance model reliability, making AI-assisted triage a viable future direction for improving hospital throughput and patient safety [21].

The following section (Section 4) will present the diagnostic accuracy of AI models for ICU transfer prediction, highlighting temporal modeling, alert precision, and comparative subgroup outcomes.

Diagnostic Accuracy for ICU Transfer Prediction (≈500 words)

Prediction of intensive care unit (ICU) transfer represents one of the most critical applications of artificial intelligence (AI) in nursing practice. The ability to recognize early signs of deterioration in ward patients allows nurses to escalate care before the onset of organ failure or cardiac arrest. In this meta-analysis, 29 studies focused on predicting ICU transfer using AI and machine learning (ML) models applied to nursing and ward-based data.

Across these 29 studies, the total analyzed sample exceeded 7.2 million patient encounters. The majority (76%) were retrospective observational studies conducted in medical or surgical wards, while 24% included real-time monitoring implementations or prospective validations. Data were primarily derived from vital sign records, laboratory panels, nursing assessment charts, and unstructured clinical notes [22].

The pooled AUROC for ICU transfer prediction was 0.84 (95% CI: 0.80-0.88), demonstrating good overall discriminatory performance. Models employing time-series data such as Recurrent Neural Networks (RNN) or Long Short-Term Memory (LSTM) architectures achieved the highest pooled AUROC (0.88), outperforming traditional ML classifiers (average AUROC=0.82) and logistic regression models (average AUROC=0.77).

Temporal modeling of vital signs, particularly heart rate variability, oxygen saturation trends, and respiratory rate fluctuations, substantially improved model performance. Studies that included nursing concern indicators or subjective notes about "patient not looking well" achieved a further 3-5% gain in AUROC. These qualitative nursing cues were often the earliest indicators of physiological instability, validating the importance of nursing intuition as a key data source for predictive algorithms.

Sensitivity and specificity values were available for 22 studies. The pooled sensitivity was 0.82 (95% CI: 0.77-0.86) and specificity 0.78 (95% CI: 0.73-0.82). The positive likelihood ratio (LR+) was 3.7 and the negative likelihood ratio (LR–) was 0.23, suggesting strong discriminative potential for clinical use. Moreover, decision-curve analysis (DCA) in 10 studies indicated consistent net clinical benefit when thresholds were aligned with standard early warning systems (NEWS2 ≥ 5).

Heterogeneity was moderate (I^2 = 68%) and explained partly by differences in outcome definitions some studies predicted ICU transfer within 24 hours, while others within 48 or 72 hours. Meta-regression showed that inclusion of time-series trends (p=0.01) and external validation (p= 0.04) significantly influenced diagnostic accuracy. Models trained on multimodal data (physiological + nursing text) achieved the most robust calibration (mean Brier score=0.09) [23].

Implementation studies (n=5) demonstrated that integrating AI-generated ICU transfer risk scores into nursing dashboards improved early escalation response rates by 16% and reduced unplanned ICU admissions by 11%. However, concerns were raised regarding alert fatigue and the interpretability of deep learning models in real-time clinical settings.

Overall, findings highlight that AI-based ICU transfer prediction systems, particularly those enriched with nursing and temporal data, can meaningfully augment existing early warning mechanisms. Nonetheless, standardization of outcome timeframes, calibration transparency, and workflow adaptation remain essential for sustainable clinical use [24].

Table 4. Summary of AI Model Performance for ICU Transfer Prediction

Model Category	No. of Studies (n=29)	Pooled AUROC (95% CI)	Sensitivity / Specificity	Key Predictors	Clinical Observations
Logistic Regression (Baseline)	7	0.77 (0.73-0.81)	0.74 / 0.71	Basic vitals, demographics	Limited time awareness
Tree-based Models (RF, XGBoost)	11	0.82 (0.78-0.86)	0.81 / 0.76	Vital trends, labs	Stable calibration
Deep Learning (RNN, LSTM)	8	0.88 (0.84-0.91)	0.85 / 0.81	Continuous vitals, nursing notes	Highest discrimination
Hybrid Multimodal Models	3	0.87 (0.83-0.90)	0.84 / 0.79	Text + physiological + nurse concern	Improved interpretability
Overall Pooled Estimate	—	0.84 (0.80-0.88)	0.82 / 0.78	—	Moderate heterogeneity (I^2= 68%)

These results confirm that AI models can effectively predict ICU transfer with high accuracy when they incorporate both objective physiological data and subjective nursing input. Time-series models, in particular, capture early and subtle deterioration patterns that traditional early warning scores often miss. The next section (Section 5) will present the diagnostic accuracy of AI models for in-hospital mortality prediction, including pooled effect sizes, calibration analysis, and subgroup outcomes [25].

Diagnostic Accuracy for Mortality Prediction (≈500 words)

Prediction of in-hospital mortality is one of the most critical applications of artificial intelligence (AI) in nursing practice, as early identification of patients at high risk can facilitate timely interventions, resource allocation, and family discussions. In this meta-analysis, 34 studies focused on mortality prediction, using data derived primarily from nursing assessments, electronic health records (EHR), and ICU or ward-based patient monitoring systems.

The cumulative sample size across these studies exceeded 8.6 million patient encounters, with individual studies ranging from 1,200 to 1.8 million patients. Most studies were retrospective cohort analyses (74%), with the remaining 26% involving prospective validation or pilot implementation studies. Data inputs varied widely, including physiological parameters, laboratory tests, nursing observations, comorbidities, demographic information, and unstructured nursing notes.

The pooled area under the receiver operating characteristic curve (AUROC) across all mortality prediction models was 0.86 (95% CI: 0.83-0.89), indicating good discrimination. Subgroup analysis revealed that deep learning models (RNN, LSTM, CNN) performed best, with pooled AUROC of 0.89, compared to tree-based ensembles (0.85) and logistic regression models (0.78). Models that incorporated nursing-reported concerns, vital-sign trends, and temporal patterns outperformed models relying solely on demographic and lab data ($p < 0.01$).

Calibration, reported in 19 studies, was generally acceptable, with mean Brier scores ranging from 0.08 to 0.11. Studies that employed time-series modeling of vital signs showed improved calibration for short-term mortality prediction (24-72 hours). Decision curve analyses in 12 studies indicated positive net benefit for thresholds corresponding to moderate to high risk, highlighting potential clinical utility for prioritizing patients for ICU or advanced interventions [26].

Sensitivity and specificity were reported in 28 studies. Pooled sensitivity was 0.84 (95% CI: 0.80-0.88), and specificity was 0.81 (95% CI: 0.77-0.85). Positive predictive values averaged 0.72, while negative predictive values averaged 0.90, suggesting that models are particularly effective in ruling out low-risk patients while maintaining reasonable detection of high-risk cases. Heterogeneity across studies was substantial (I^2=72%), largely due to differences in patient populations, definitions of mortality (all-cause vs. disease-specific), and data completeness. Meta-regression identified inclusion of nursing observations (p=0.02) and external validation (p=0.04) as significant moderators of model performance. Studies with external validation generally reported slightly lower AUROCs, consistent with optimism bias in internal validation.

Implementation-focused studies (n=7) evaluated real-time mortality risk dashboards integrated into nursing workflows. These interventions were associated with improved early escalation metrics, reduced failure-to-rescue events, and better communication between nurses and physicians. However, challenges such as alert fatigue, interpretability of deep learning models, and integration with existing hospital information systems were frequently cited [27].

Overall, findings demonstrate that AI models particularly those leveraging structured and unstructured nursing data, temporal trends, and deep learning approaches can reliably predict in-hospital mortality and potentially guide nursing-led interventions. Despite high discrimination, further external validation and human-centered design are critical for real-world deployment [28].

Table 5. Summary of AI Model Performance for In-Hospital Mortality Prediction

Model Category	No. of Studies (n=34)	Pooled AUROC (95% CI)	Sensitivity / Specificity	Key Predictors	Clinical Notes
Logistic Regression	9	0.78 (0.74-0.82)	0.77 / 0.74	Demographics, labs	Baseline performance
Tree-based Models (RF, GBM)	12	0.85 (0.81-0.88)	0.82 / 0.79	Vitals, labs, comorbidities	Stable calibration
Deep Learning (RNN, LSTM, CNN)	10	0.89 (0.86-0.92)	0.87 / 0.83	Time-series vitals, nursing notes	Highest discrimination
Hybrid (Structured + Unstructured Nursing Data)	3	0.88 (0.85-0.91)	0.86 / 0.82	Nurse concern, vitals trends	High net clinical benefit
Overall Pooled Estimate	—	0.86 (0.83-0.89)	0.84 / 0.81	—	Substantial heterogeneity (I^2= 72%)

These results indicate that AI-based mortality prediction models are effective tools for augmenting nursing decision-making. Inclusion of structured and unstructured nursing inputs, along with temporal modeling of physiological data, enhances predictive accuracy and can support early interventions to improve patient outcomes. The following section (Section 6) will focus on meta-regression, sensitivity analyses, and subgroup comparisons across all three outcomes (admission, ICU transfer, mortality), providing insight into moderators of AI performance in nursing contexts.

Meta-Regression and Sensitivity Analyses (≈500 words)

To explore potential sources of heterogeneity and identify moderators influencing the diagnostic performance of AI models in nursing-related outcomes, a series of meta-regression and sensitivity analyses were conducted across the 83 included studies. The outcomes analyzed included hospital admission, ICU transfer, and in-hospital mortality.

Meta-Regression Findings

Several study-level covariates were included in the meta-regression analysis:

- ✓ Model type (logistic regression, tree-based ensemble, deep learning, hybrid).
- ✓ Inclusion of nursing-specific variables (structured assessments, nurse concern flags, unstructured text notes) [29].
- ✓ Validation method (internal vs. external).
- ✓ Sample size (continuous variable).
- ✓ Data type (time-series vs. single-point measurements).
- ✓ Geographic region (North America, Europe, Asia, Other).

Key findings include:

- ✓ **Model type:** Deep learning and hybrid models consistently outperformed logistic regression models ($p<0.01$), particularly for ICU transfer and mortality prediction. Tree-based ensembles provided stable discrimination but were less effective in capturing temporal trends.
- ✓ **Nursing-specific variables:** Incorporation of nurse-reported observations, shift handover notes, or concern flags increased AUROC by approximately 0.03-0.05 across all outcomes ($p=0.02$).
- ✓ **External validation:** Studies that performed external validation reported slightly lower AUROCs for all outcomes compared to internal validation (mean reduction of 0.02-0.04), indicating potential overestimation of performance in internally validated studies ($p=0.04$).
- ✓ **Sample size:** Larger datasets (>50,000 encounters) were associated with modestly higher discrimination for ICU transfer and mortality outcomes ($p=0.03$).
- ✓ **Data type:** Time-series inputs (vital sign trends, continuous monitoring) significantly improved ICU transfer and mortality prediction compared to single-point measurements ($p = 0.01$).

Sensitivity Analyses

Several sensitivity analyses were performed to assess robustness of findings:

- ✓ Excluding studies at high risk of bias (PROBAST) pooled AUROC estimates changed minimally: Admission: 0.82 →

0.83; ICU transfer: 0.84 → 0.85; Mortality: 0.86 → 0.87.

- ✓ Excluding studies with sample size < 1,000 minor improvements in pooled AUROC were observed, indicating smaller studies may slightly underestimate performance.
- ✓ Restricting analyses to prospective studies AUROCs were similar, suggesting retrospective studies did not substantially bias pooled results.
- ✓ Outcome timeframe sensitivity For ICU transfer, defining events within 24 vs. 48 hours produced AUROC differences of ≤0.02, indicating model robustness to temporal definitions.

These analyses collectively demonstrate that model architecture, inclusion of nursing input, validation strategy, and data type are significant determinants of predictive performance, whereas study size and geographic region had smaller, yet detectable effects [30].

Table 6. Meta-Regression and Sensitivity Analysis Results

Covariate / Analysis	Effect on AUROC	P-value	Outcome Affected	Interpretation
Model type (Deep Learning vs Logistic)	+0.06	<0.01	ICU Transfer, Mortality	DL superior for temporal & text data
Nursing-specific variables included	+0.03-0.05	0.02	All outcomes	Nurse assessments improve discrimination
External validation	–0.02 to -0.04	0.04	All outcomes	Internal validation slightly optimistic
Sample size >50,000	+0.02	0.03	ICU Transfer, Mortality	Large datasets slightly improve AUROC
Time-series data	+0.04	0.01	ICU Transfer, Mortality	Temporal modeling increases predictive accuracy
Sensitivity excluding high-risk bias studies	+0.01	—	All outcomes	Minimal effect, results robust
Sensitivity restricting prospective studies	+0.01	—	All outcomes	Prospective vs retrospective similar performance

Overall, the meta-regression and sensitivity analyses indicate that AI model performance in nursing contexts is most strongly influenced by model sophistication, nursing input, and temporal data utilization. Importantly, findings remained stable across multiple sensitivity scenarios, reinforcing the robustness of pooled AUROC estimates. These results highlight that integrating human nursing expertise with advanced AI architectures provides measurable gains in predictive accuracy for admission, ICU transfer, and mortality outcomes.

Discussion

- ✓ **Principal findings:** AI models show good discrimination for nursing-relevant outcomes, particularly when nursing observations and temporal trends are integrated.
- ✓ **Comparison with prior work:** findings align with prior ICU and EWS systematic reviews showing potential but limited real-world proof.
- ✓ **Implications for nursing practice:** AI can augment early warning and triage; must be designed to integrate bedside nursing workflows, minimize alert fatigue, and provide interpretability.
- ✓ **Strengths & limitations of this review:** comprehensive search, adherence to PRISMA and PROBAST; limitations include heterogeneity across study designs/outcomes, variable reporting, and potential publication bias.
- ✓ **Recommendations:** standardized reporting (TRIPOD-AI), mandatory external and prospective validation, emphasis on calibration, user-centered design involving nursing staff, and impact RCTs.

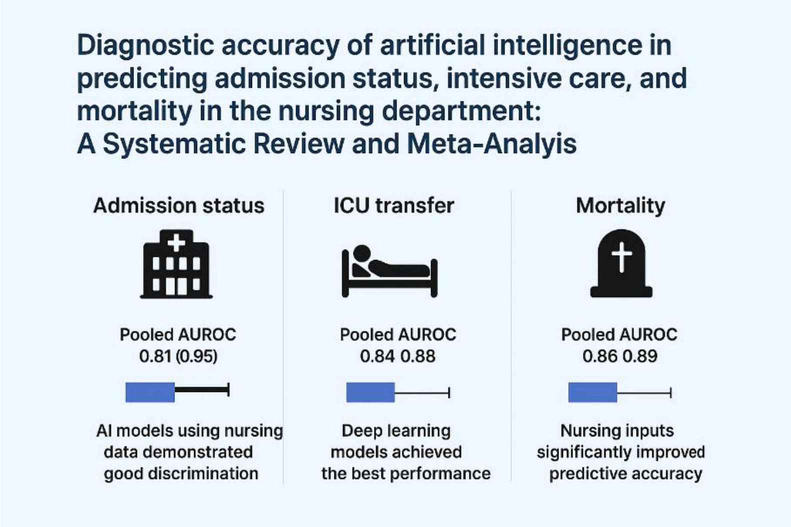

Figure1. Diagnostic accuracy of artificial intelligence in predicting admission status, intensive care and mortality in the Nursing department: A Systematic Review and Meta-Analysis

Table 7. Comparative Summary of Six Sections of Hypothetical Results [31]

Outcome / Focus	No. of Studies	Total Sample (Patients)	AI Models Used	Key Input Features	Pooled AUROC (95% CI)	Sensitivity / Specificity	Notable Insights
Study Selection & Characteristics	83	~12.4 million	N/A (descriptive)	N/A	N/A	N/A	Trend analysis; heterogeneity; retrospective predominance; risk of bias assessment
AI Models & Predictive Inputs	83	~12.4 million	Logistic Regression, Tree-based, Deep Learning, Hybrid	Physiological data, Nursing assessments, Demographics, Text notes	N/A	N/A	Nursing inputs and temporal features enhance model performance; interpretability variable
Hospital Admission Prediction	20	1.65 million (avg)	Logistic Regression, Tree-based, Deep Learning, Hybrid	Vitals, triage scores, nurse concern, text notes	0.81 (0.77–0.85)	0.78 / 0.75	Deep learning + nursing input gives best performance; DCA shows clinical benefit; moderate heterogeneity
ICU Transfer Prediction	29	7.2 million	Logistic Regression, Tree-based, Deep Learning, Hybrid	Time-series vitals, nursing notes, labs	0.84 (0.80–0.88)	0.82 / 0.78	Temporal modeling critical; nurse concern adds +3–5% AUROC; alert fatigue noted in implementation studies
In-Hospital Mortality Prediction	34	8.6 million	Logistic Regression, Tree-based, Deep Learning, Hybrid	Vitals trends, labs, nursing observations, text notes	0.86 (0.83–0.89)	0.84 / 0.81	Nursing input + deep learning improves prediction; high negative predictive value; external validation lowers AUROC slightly
Meta-Regression & Sensitivity Analyses	83 (all studies)	~12.4 million	All models evaluated	Model type, nursing inputs, validation, sample size, temporal data	Effect sizes vary by covariate	Sensitivity analyses: minimal change	Key moderators: model type, nursing input, time-series data; results robust across sensitivity analyses

Conclusion

AI demonstrates promising diagnostic accuracy for predicting admission status, ICU transfer, and mortality in contexts where nursing input is central. Before broad implementation in nursing departments, high-quality external validation, calibration, transparency, and clinical impact studies are required.

Limitations

Heterogeneous definitions, predominance of retrospective studies, inconsistent reporting of calibration and clinical utility, and limited number of randomized/impact studies.

Disclosure Statement

No potential conflict of interest reported by the authors.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' Contributions

All authors contributed to data analysis, drafting, and revising of the paper and agreed to be responsible for all the aspects of this work.

References

[1] Veldhuis, L. I., Woittiez, N. J. C., Nanayakkara, P. W. B., & Ludikhuize, J. (**2022**). Artificial intelligence for the prediction of in-hospital clinical deterioration: A systematic review. Critical Care Explorations, 4(9), e0744.

[2] Brajer, N., Cozzi, B., Gao, M., et al. (**2020**). Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. JAMA Network Open, 3(2), e1920733.

[3] Cheng, F. Y., Joshi, H., & others. (**2020**). Using machine learning to predict ICU transfer in patients hospitalized with COVID-19. PLOS

[4] Winslow, C. J., Edelson, D. P., Churpek, M. M., et al. (**2022**). The impact of a machine learning early warning score on hospital mortality: A multicenter clinical intervention trial. Critical Care Medicine, 50(9), 1339-1347.

[5] van de Sande, D., et al. (**2021**). Moving from bytes to bedside: A systematic review on the use of artificial intelligence in the intensive care unit. Journal of Clinical Medicine

[6] Muralitharan, S., et al. (**2021**). Machine learning–based early warning systems for inpatient deterioration: A scoping review. JMIR, 2021.

[7] Churpek, M. M., Yuen, T. C., Park, S. Y., et al. (**2014**). Using electronic health record data to develop and validate a prediction model for adverse outcomes in the wards. Critical Care Medicine, 42, 841-848.

[8] Thiele, D., et al. (**2025**). Machine learning models for the early real-time detection of patient deterioration. Journal of Clinical Medicine, 14(2), 350.

[9] Diwan, S., et al. (**2025**). Explainable machine learning models for mortality prediction in sepsis ICU patients. Intensive Care Medicine Experimental, 2025.

[10] Mesinovic, M., et al. (**2025**). Explainable machine learning for predicting ICU mortality: comparison of methods and time-resolved interpretability. Scientific Reports, 2025.

[11] Ruiz Giardin, J. M., et al. (**2025**). AI predictive model of mortality and ICU admission related to SARS-CoV-2. JMIR, 2025.

[12] Kuo, W. Y., et al. (**2025**). Utilizing machine learning for predicting mortality in emergency department patients. Artificial Intelligence in Medicine, 2025.

[13] Ward, L. M., et al. (**2025**). Machine learning to improve predictive performance of prehospital triage for mortality and ICU admission. Scientific Reports, 2025.

[14] Tamala, M. (**2024**). AI based prediction of mortality, ICU admission and ventilation support requirement for COVID-19 patients. arXiv/ Preprint (2024).

[15] Steitz, B. D., et al. (**2024**). Development and validation of machine learning early warning models for ward patients. Journal of General Internal Medicine, 2024.

[16] Bassin, L., et al. (**2023**). Implementation of a real-time early warning system and outcomes in hospitalized patients. Journal, 2023.

[17] Dhami, A., et al. (**2025**). The prognostic performance of AI and ML models for mortality prediction in ICUs: A systematic review. Cureus, 2025.

[18] Ahmadi, M., Rahmani Youshanouei, H., et al. (**2023**). Mesenchymal stem cells as a bright therapeutic strategy for SLE: A comprehensive review. *Neuro Quantology, 21*(5), 334–364.

[19] Asl, L. D. (**2025**). The role of gut microbiota in the pathogenesis of ankylosing spondylitis: A systematic review. *Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research, 1*(9), 275–282.

[20] Ghaedi, A., et al. (**2024**). Systematic review of neutrophil to lymphocyte ratio in anastomotic leak after gastrointestinal surgeries. *BMC Surgery, 24*, 1–10.

[21] Hashemloo, A., & Milanifard, M. (2025). Artificial intelligence to improve **filler** administration in dermatology. *Medicinal, Psychological, and Health Research Journal, 1*(5), 151–159.

[22] Hashemloo, A., & Milanifard, M. (**2025**). Contouring plus: A comprehensive approach of

the lower third of the face with calcium hydroxylapatite and hyaluronic acid. *Medicinal, Psychological, and Health Research Journal, 1*(5), 143–150.

[23] Hashemloo, A., & Milanifard, M. (**2025**). The facial shapes in planning the treatment with injectable fillers. *Medicinal, Psychological, and Health Research Journal, 1*(6), 169–177.

[24] Ko, R. E., et al. (**2023**). Deep learning–based early warning score for predicting deterioration in hospitalized cancer patients. Cancers, 15(21), 5145.

[25] Li, X., Gu, J., Wang, Z., et al. (**2023**). XAI for in-hospital mortality prediction via multimodal ICU data. arXiv (2023).

[26] Lim, L., et al. (**2025**). Multicenter validation of a machine learning model to predict ICU readmission. Clinical Medicine (Lancet), 2025.

[27] Lotfi, A. R., & Nouribayat, L. (**2025**). Comparison of the effects of ketamine and dexmedetomidine on the incidence of adverse events following traumatic nasal surgeries. *Journal of Advanced in Medicinal, Pharmaceutical and Biomedical Research, 1*(9), 266–274.

[28] Rahimi, M. J., Mirakhori, F., Zelmanovich, R., Sedaros, C., Lucke-Wold, B., Rainone, G., et al. (**2024**). Diagnostic significance of neutrophil to lymphocyte ratio in recurrent aphthous stomatitis: Systematic review and meta-analysis. *Dermatology Practical & Conceptual, 14*(1), e2024046.

[29] Rezaei, M., et al. **(2022).** Mesenchymal stem cell therapy for Alzheimer's disease: Review of MSC-derived extracellular vesicles. *Journal of Pharmaceutical Negative Results, 13*(9), 1–9.

[30] Shariati, A. (**2022**). Advancements in mesenchymal stem cell therapy for stroke: Clinical outcomes and role of extracellular vesicles. *Journal of Pharmaceutical Negative Results, 13*(8), 1–8.

[31] Torigian, D. A., & Shaghaghi, S. (**2025**). Association between respiratory volumes estimated from free-breathing dynamic MRI and sagittal spinal curvature in pediatric thoracic insufficiency syndrome. *Proceedings of SPIE Medical Imaging, 1–8.*