ACHSM Asia-Pacific Journal of Health Management

REVIEW ARTICLE

ARTIFICIAL INTELLIGENCE BASED PREDICTION OF EMERGENCY DEPARTMENT REVISITS AND HOSPITALIZATION: A SYSTEMATIC REVIEW AND META-ANALYSIS

Roxana Hessam[1], Seyed Mohammad Hosseini Kasnavieh[2], Seyed Hossien Shaker[3], Ali Saghandian Tousi[1], Maryam Milanifard[4,5]*

1. Department of Emergency medicine, school of medicine, Iran University of medical sciences, Tehran, Iran
2. Associate professor of Emergency medicine, Department of Emergency medicine, school of medicine, Hazrat-e Rasool General Hospital, Iran University of medical sciences, Tehran, Iran
3. Assistant Professor of Internal Medicine, Department of Emergency Medicine, School of Medicine, Hazrat-e Rasool General Hospital, Iran University of Medical Sciences, Tehran, Iran
4. Trauma and injury Research center, Iran University of medical sciences, Tehran, Iran
5. Student Research committee, Iran University of medical sciences, Tehran, Iran

*Correspondence: mhoseini1346@mailfa.com

ABSTRACT

BACKGROUND AND AIM:

Short-term emergency department readmission is one of the major challenges in resource management and quality of clinical care. The aim of this study was to evaluate the performance of artificial intelligence models in predicting patients' frequent emergency department visits and to investigate the effect of the data type and algorithm used on the sensitivity and specificity of prediction.

METHOD:

The international databases PubMed, EMBASE, and Web of Science were searched using keywords aligned with the study objective between from January 2015 to 10 December 2025. In the present study, 27 studies were included in the meta-analysis. Data analysis was performed using the stata. v17 software and analysis was performed using a random effects model (DerSimonian–Laird) to determine sensitivity and specificity.

RESULT:

The overall sensitivity of the models was 0.61 [95% CI; 0.54 to 0.68] and the overall specificity was 0.83 [95% CI; 0.78–0.89]. Subgroup analysis showed that the type of AI algorithm had a significant effect on diagnostic accuracy; Ensemble and XGBoost models performed better.

CONCLUSION:

Artificial intelligence models can be a useful tool for identifying high-risk ED patients, especially using multi-source data and hybrid algorithms.

KEYWORDS

Artificial Intelligence, Emergency Department, Diagnostic accuracy, Revisit, hospitalization.

INTRODUCTION

Unnecessary patient readmissions and hospitalizations following emergency department (ED) visits represent one of the most significant challenges facing healthcare systems worldwide[1]. According to global statistics, about 15 to 20 percent of patients return within a month after being discharged from the ED, and more than ten percent require re-hospitalization[2, 3]. Repeated visits impose significant direct and indirect costs on the healthcare system, and studies have also reported an increased risk of hospital complications, reduced quality of patient care, and reduced patient satisfaction[4]. Studies have shown that more than 50 million ED readmissions occur annually in the United States, approximately 15 to 20 percent of which result in readmissions within 30 days, and the direct cost of these visits is estimated to be $3 to $5 billion[5]. Identifying high-risk patients at the first visit to the ED to prevent readmissions can reduce the burden on healthcare systems, and accurate prediction of these events allows for the design of preventive interventions and evidence-based clinical decision-making[6, 7]. These challenges highlight the need for more advanced predictive approaches beyond traditional statistical methods.

Over the past decade, the use of artificial intelligence (AI) has emerged as an advanced tool in predicting clinical outcomes and managing big data[8, 9]. Evidence has shown that models based on machine learning and neural networks have the ability to extract complex patterns and accurately predict emergency department revisits and hospitalizations[10-12]. Systematic reviews have highlighted the increasing use of mobile and AI-based applications in supporting decision-making processes and vulnerable populations, highlighting the scalability and adaptability of AI-based models across diverse real-world settings in diverse real-world contexts[13]. In addition to clinical applications, recent evidence from higher education has shown that AI systems can achieve acceptable levels of reliability, validity, and fairness in complex assessment tasks, although concerns regarding alignment with human judgment remain[14].

The importance of using artificial intelligence models in predicting and reducing unnecessary readmissions can reduce the economic burden, decrease hospital complications, improve healthcare quality, and enhance satisfaction among patients and emergency department staff[15]. Therefore, the development and application of artificial intelligence can have a significant impact on the efficiency, safety, and economics of healthcare systems[16, 17]. Despite the growing number of studies investigating AI-based prediction models in emergency care, the reported diagnostic performance of these models remains heterogeneous, and the influence of algorithm type and data source on predictive accuracy has not been comprehensively synthesized. Therefore, this systematic review and meta-analysis aimed to evaluate the diagnostic performance of AI-based models for predicting emergency department revisits and hospitalizations and to examine the impact of algorithm type and data source on model accuracy.

METHOD

STUDY DESIGN AND REPORTING STANDARDS

The present study was designed as a systematic review and meta-analysis to evaluate the performance of artificial intelligence models in predicting ED readmissions and hospitalization. The reporting process was conducted according to the PRISMA 2020 guidelines[18]. As this study was based exclusively on previously published data, ethical approval was not required. Ethical considerations were addressed according to the requirements reported in the original studies. All steps, including search, screening, data extraction, and analysis, were performed independently by two researchers, and disagreements were resolved with a third reviewer. The international databases PubMed, EMBASE, and Web of Science were evaluated for articles using keywords from January 2015 to 10 December 2025. In addition, a search for grey literature sources in Google Scholar and a review of the reference lists of the included studies were also conducted. The research question was defined based on the PICO framework (population, intervention, comparison, and outcome). The study protocol was developed at the beginning to clarify the purpose, inclusion and exclusion criteria, evaluation method, and data analysis method.

MESH TERMS USED TO RETRIEVE LITERATURES:

((((((((((((((((("Emergency Service, Hospital"[Mesh] OR "Emergency Room Visits"[Mesh]) OR "Emergency Service, Hospital"[Majr]) OR "Emergency Room Visits"[Majr]) OR "Emergency Medical Services"[Mesh]) AND ("Patient Readmission"[Mesh] OR "Patient Admission"[Mesh])) OR "Patient Readmission"[Majr]) OR "Hospitalization"[Mesh]) OR "Hospitalization"[Majr]) OR "Patient Admission"[Majr]) AND "Artificial Intelligence"[Mesh]) OR "Artificial Intelligence"[Majr]) OR ("Artificial Intelligence/classification"[Majr] OR "Artificial Intelligence/statistics and numerical data"[Majr])) OR "Machine Learning"[Mesh]) OR "Deep Learning"[Mesh]) AND "Data Accuracy"[Mesh]) OR "Area Under Curve"[Mesh]) AND ("Sensitivity and Specificity"[Mesh] OR "Parasitic Sensitivity Tests"[Mesh])) AND "Large Language Models"[Mesh]) AND "Diagnosis"[Mesh].

STUDY INCLUSION AND EXCLUSION CRITERIA

Studies that met the requirements of the PICO strategy were included:

Population (P): Patients presenting to the ED without age or diagnostic restrictions.

Intervention (I): Artificial intelligence

Comparison (C): Traditional method/lack of comparator.

Outcome (O): Diagnostic accuracy, sensitivity, specificity

Animal studies, use of non-algorithmic methods, review articles, letters to the editor, conferences, duplicate articles, studies published in languages other than English, failure to provide complete data, and studies that did not have specific objectives were excluded from the study.

DATA EXTRACTION

Data from each included study were independently extracted by two investigators using a standard data extraction form; disagreements between the investigators were discussed and settled by a third researcher. The following data were extracted from each study: first author, publication year, sample size, revisit interval, data source, and AI technique.

QUALITY ASSESSMENT

In the present study, the Predictive Model Bias Assessment Tool (PROBAST) was used to assess the risk of bias in predictive model studies. This tool has 20 questions in four areas (Participants, Predictors, Outcome, and Analysis)[19].

STATISTICAL ANALYSIS

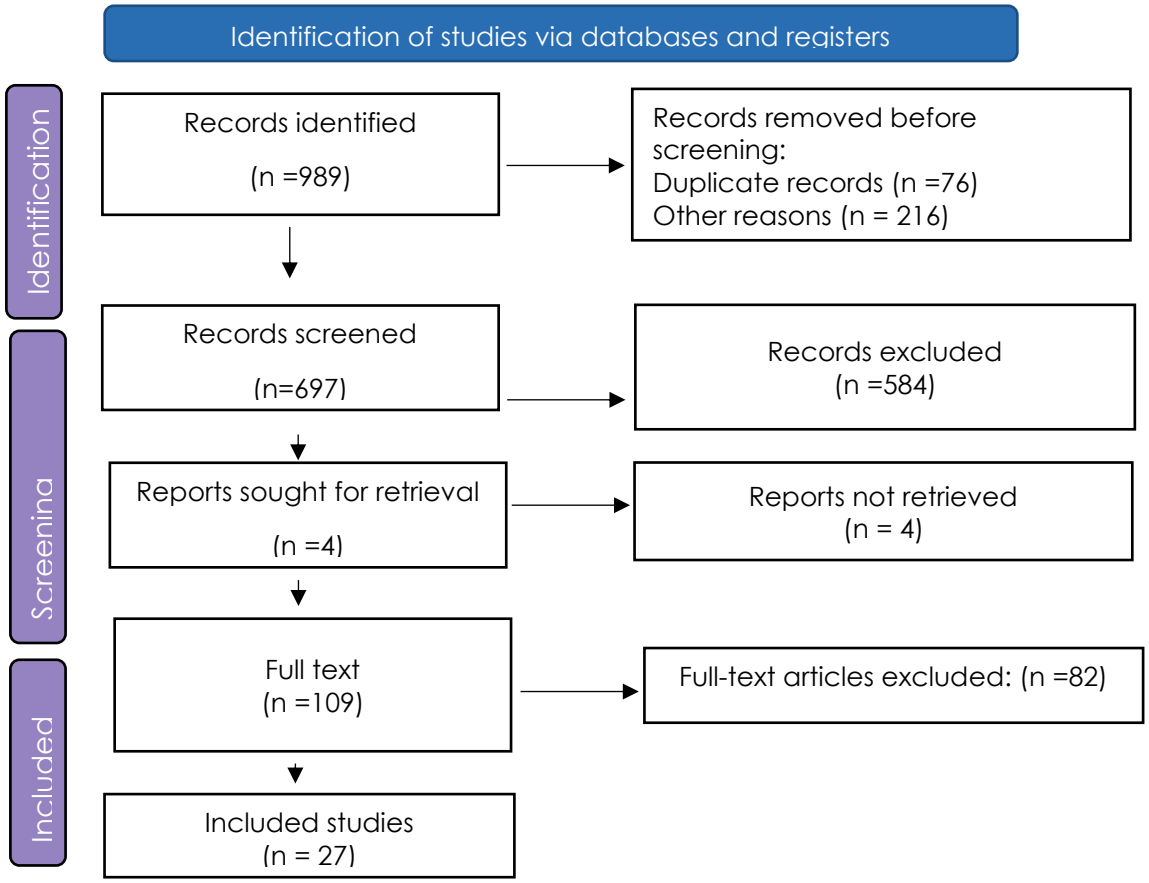
Data analysis was performed using STATA/MP.v17 (College Station, Texas, USA). $p<0.05$ was considered as statistically significant difference. Pooled sensitivity and specificity were calculated using a random-effects model based on the DerSimonian–Laird method. Statistical heterogeneity was assessed using Cochran's Q test and the I^2 statistic: No heterogeneity: $0.0\% < I^2 < 24.9\%$; low heterogeneity: $25.0\% < I^2 < 49.9\%$; Moderate heterogeneity: $50.0\% < I^2 < 74.9\%$; High heterogeneity: $75.0\% < I^2 < 100\%$.

RESULT

LITERATURE SEARCH

The database search identified 989 records, as shown in Figure 1 . After reviewing the titles, 292 of these articles were excluded based on duplication and inconsistency with the inclusion and exclusion criteria; the abstracts of 697 articles were reviewed and studies that did not meet the study selection criteria were excluded at this stage (n=584), and the full text of 109 articles was reviewed by two blind and independent authors. Finally, after removing irrelevant articles, 27 articles were included in the meta-analysis.

FIGURE 1. PRISMA 2020 FLOW DIAGRAM



STUDY CHARACTERISTICS

In the present study, the total sample size was 3,052,833 patients, which included a wide range of data based on electronic health records (EHR/EMR), insurance databases (NHIRD), and multicenter data. Revisit intervals varied substantially across studies. In most studies, 72 hours have been considered, while in others, up to 9 days, 30 days, 120 hours, 7 days, and even three months have been reported. Overall, the heterogeneous nature of the time interval for re-visits, the diversity of clinical characteristics, and the differences in algorithms used across studies highlight the need for a comprehensive meta-analysis to combine results and provide an accurate estimate of the performance of AI models. Collectively, the included studies provided data from more than three million patients, enabling robust quantitative synthesis. Table 1 shows a summary of the study data.

BIAS ASSESSMENT

Of the 27 studies, 14 were assessed as low risk of bias and 13 as having a moderate risk of bias. Most studies had a well-defined population or sufficient sample size, but some had incomplete reporting, which made their status unclear. (Table 2).

TABLE 1. SUMMARY OF DATA EXTRACTED FROM STUDIES

No.	Study. Years	Number of patients	ED revisit interval (Hour)	AUC	Source	AI model
1	Avendano et al., 2025 [20]	3591	72, 216, 720	0.85	EHR	ML
2	Oh et al., 2025 [21]	12977	72	0.74	EHR	ML
3	Liu et al., 2025 [22]	38797	72	0.81	CETAT	MLP

4	Bians et al., 2025 [23]	8969	72	0.84	EHR	ML
5	Sadek et al., 2025 [24]	2023	72	0.75	EHR	XGBoost
6	Watson et al., 2024 [25]	174393	72	0.84	EHR	ML
7	Cohen et al., 2024 [26]	213	3 months	0.71	EHR	ML
8	Hsu et al., 2024 [27]	25225	72	0.71	EHR	Voting
9	Chen et al., 2024 [28]	128848	72	0.64	EHR	BERT
10	Zhang et al., 2024 [29]	62154	72	0.87	EHR	LR
11	Sung et al., 2024 [30]	6282	72	0.82	EMR	stacking Ensemble
12	Saggu et al., 2024 [31]	4473	720	0.70	EHR	GNN
13	Davazdahemami et al., 2022 [32]	657752	72	0.74	EHR	DNN
14	Aldhoayan et al., 2022 [33]	77906	168	0.61	EHR	CatBoost
15	Li et al., 2021 [34]	436810	72	0.93	EHR	XGBoost
16	Hsu et al., 2021 [35]	25151	72	0.74	EHR	Voting
17	Chmiel et al., 2021 [36]	44294	72	0.74	EHR	XGBoost
18	Chen et al., 2021 [37]	49252	72	0.76	NHIRD	C4.5
19	Artetxe et al., 2020 [38]	99858	72	0.64	EHR	Bagging
20	Vest et al., 2019 [39]	279611	720	0.73	EHR	DT
21	Meng et al., 2019 [40]	328733	72	0.73	EMR	DA
22	Hong et al., 2019 [41]	330631	72	0.76	EHR	XGBoost
23	Bergese et al., 2019 [42]	28341	120	0.81	EHR	CT
24	Hu et al., 2017 [43]	125940	120	0.72	EMR	RF
25	Artetxe et al., 2017 [44]	99858	72	0.60	EMR	RF
26	Artetxe et al., 2016 [45]	360	720	0.88	EMR	SVM
27	Besga et al., 2015 [46]	391	720	0.72	EHR	SVM

ED = Emergency Department; EHR = Electronic Health Records; EMR = Electronic Medical Records; CETAT = Clinical Emergency Triage Assessment Tool; DNN = Deep Neural Network; SVM = Support Vector Machine; NHIRD = National Health Insurance Research Database.

TABLE 2. RISK OF BIAS OF PREDICTION MODEL RISK OF BIAS ASSESSMENT TOOL

Study. Years	Participants	Predictors	Outcome	Analysis	Overall
Avendano et al., 2025 [20]	L	L	L	L	Low
Oh et al., 2025 [21]	L	L	L	L	Low
Liu et al., 2025 [22]	L	L	L	L	Low
Bians et al., 2025 [23]	L	L	L	L	Low
Sadek et al., 2025 [24]	L	L	L	L	Low

Watson et al., 2024 [25]	L	L	L	L	Low
Cohen et al., 2024 [26]	L	L	L	L	Low
Hsu et al., 2024 [27]	U	L	L	L	Moderate
Chen et al., 2024 [28]	U	L	L	L	Moderate
Zhang et al., 2024 [29]	L	L	L	L	Low
Sung et al., 2024 [30]	U	L	L	L	Moderate
Saggu et al., 2024 [31]	U	L	L	L	Moderate
Davazdahemami et al., 2022 [32]	U	L	L	L	Moderate
Aldhoayan et al., 2022 [33]	L	L	L	L	Low
Li et al., 2021 [34]	L	L	L	L	Low
Hsu et al., 2021 [35]	L	L	L	L	Low
Chmiel et al., 2021 [36]	L	L	L	L	Low
Chen et al., 2021 [37]	U	L	L	L	Moderate
Artetxe et al., 2020 [38]	U	L	L	L	Moderate
Vest et al., 2019 [39]	L	L	L	L	Low
Meng et al., 2019 [40]	L	L	L	L	Low
Hong et al., 2019 [41]	L	L	L	L	Low
Bergese et al., 2019 [42]	L	L	L	L	Low
Hu et al., 2017 [43]	U	L	L	L	Moderate
Artetxe et al., 2017 [44]	U	L	L	L	Moderate
Artetxe et al., 2016 [45]	L	L	L	L	Low
Besga et al., 2015 [46]	U	L	L	L	Moderate

L: Low risk of bias; U: Unclear risk of bias; H: High risk of bias

SENSITIVITY OF ED REVISIT

The overall sensitivity of the AI models was reported to be 0.60 with a 95% confidence interval between 0.52 and 0.68 (Figure 2). This value indicates that the predictive AI models were able to correctly identify about 60% of actual patient returns to the ED. This sensitivity level is in the medium range, indicating that a significant percentage of repeat visits are still not identified by artificial intelligence models and that there is a need to optimize and improve the algorithms.

Subgroup analysis was performed to investigate the differences in performance of AI models based on the Revisit timeframe. The results showed that the sensitivity was different in different groups, such that after 72 hours the sensitivity was estimated to be 0.60 (95% CI: 0.52–0.70) with very high heterogeneity (I^2 = 89.4%). Revisit timeframe after 30 days was estimated with a sensitivity of 0.59 (95% CI: 0.46–0.71) and high heterogeneity (I^2 = 79.3%), which showed the performance of AI models slightly lower than the 72-hour prediction. Sensitivity was estimated to be 0.66 and 0.91 after 7 days and 3 months, respectively. The AI prediction of readmission after 3 months showed high performance, but its generalizability is limited due to being a single study. The sensitivity after nine days was estimated to be 0.438 (95% CI: 0.046–0.830) with very high heterogeneity (I^2 = 97.3%). The test of difference between groups (Q_b = 29.30, $p < 0.001$) showed that there was a significant difference between subgroups.

The overall heterogeneity of the studies was very high (I^2 = 91.9%). This high level of heterogeneity is due to differences between studies; type of AI algorithms, sample size, length of patient return interval, data quality and study design.

SPECIFICITY OF ED REVISIT

The overall specificity was reported to be 0.84 with a 95% confidence interval between 0.79 and 0.90, the AI models were able to correctly identify more than 84% of non- revisits patients, indicating a relatively good performance of the AI models (Figure 3).

FIGURE 2. FOREST PLOT SHOWED SENSITIVITY OF PREDICTING ED REVISIT

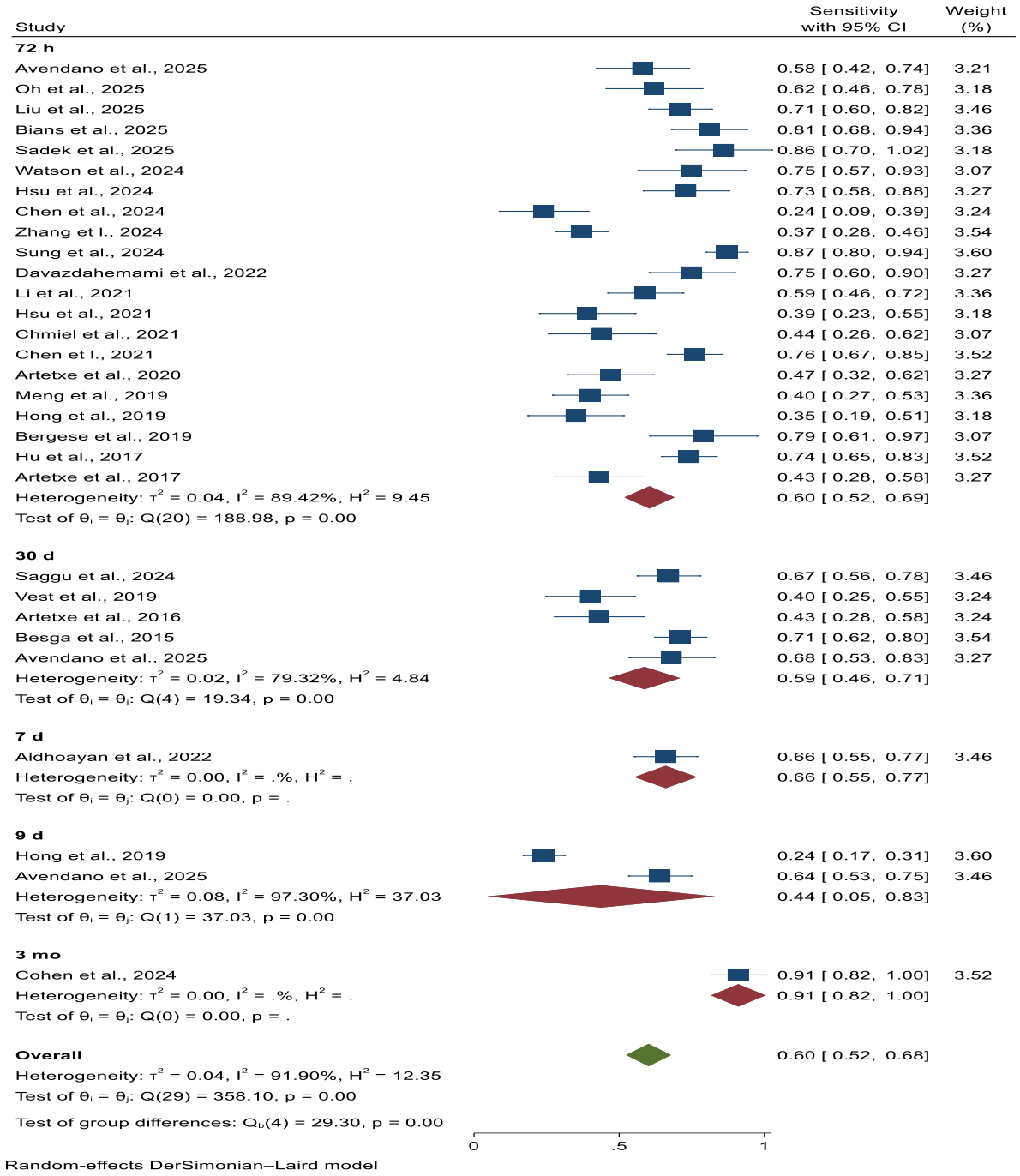
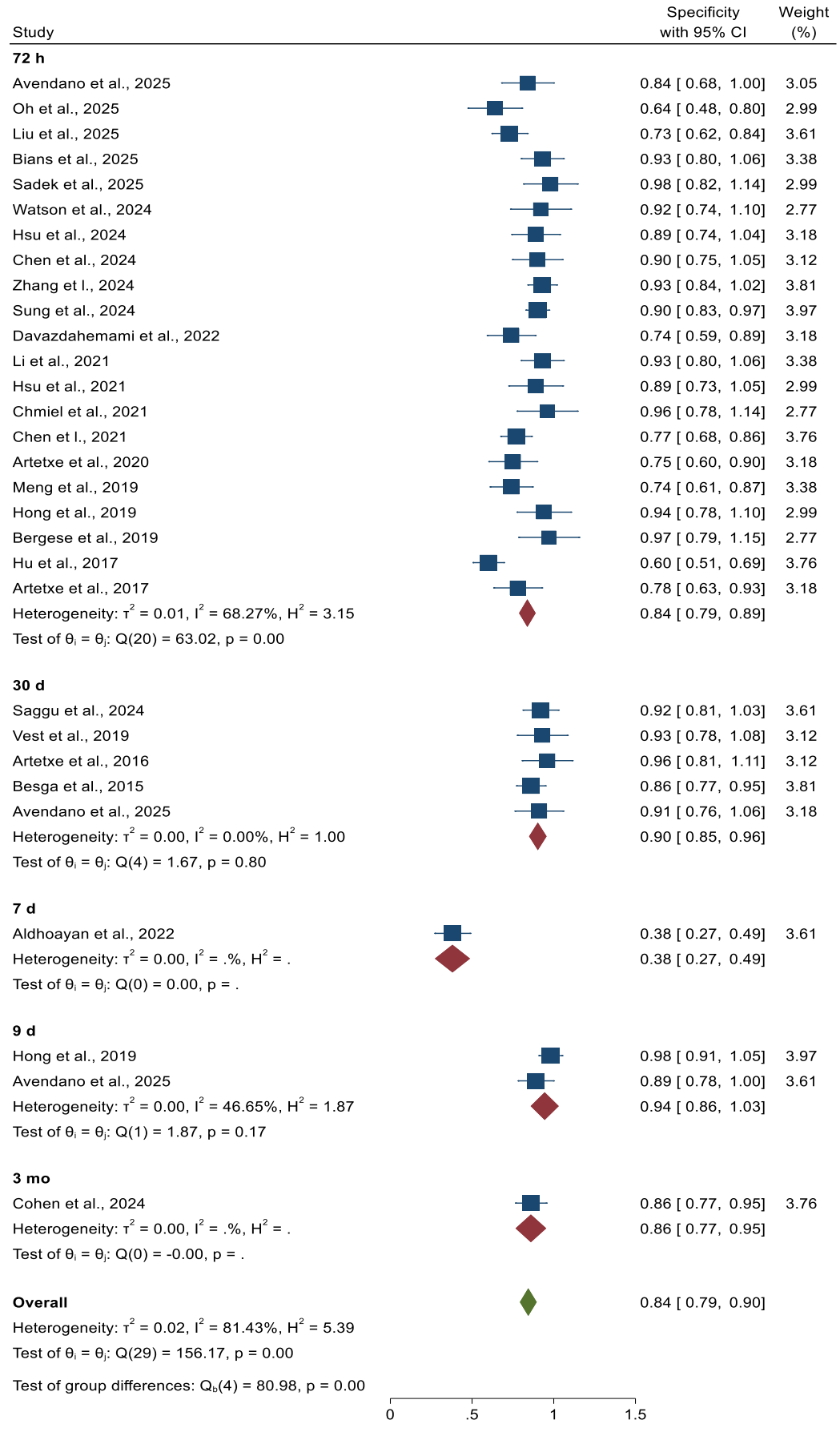


FIGURE 3. FOREST PLOT SHOWED SPECIFICITY OF PREDICTING ED REVISIT



Subgroup meta-analysis was conducted to determine the performance of AI models based on the Revisit timeframe (Figure 3). After 72 hours, the specificity was estimated to be 0.84(95% CI: 0.79–0.89) with high heterogeneity (I^2 = 68.3%). After 30 days, specificity was 0.90 (95% CI: 0.85–0.96) with very low heterogeneity (I^2 = 0.0%). This group had consistent performance, indicating that the models in this group were able to correctly identify approximately 90% of non- revisits patients. After 7 days, the specificity was 0.38, which is relatively low and indicates poor performance of the AI model. After 9 days, specificity was estimated to be 0.95 (95% CI: 0.86–1.031) with moderate heterogeneity (I^2 = 46.7%), indicating very good performance of the models in this group. Specificity was also estimated to be 0.86 after three months, indicating good and reliable performance of the AI model. The difference between groups test (Q_b = 80.98, $p < 0.001$) showed that there were significant differences between subgroups and that the performance of AI models differed based on Revisit timeframe. Overall heterogeneity between studies was high (I^2 = 81.4%), indicating significant differences between algorithms, sample size, patient return timeframe, and data quality.

Due to the high heterogeneity between studies, a subgroup analysis of meta-analysis was performed based on data source and type of artificial intelligence algorithm (Figure 4, 5 and Table 3, 4). Subgroup meta-analysis by data source showed that there was no statistically significant difference between subgroups (Qb = 6.82, p = 0.08) (Figure 4). The highest sensitivity was for national structured databases such as NHIRD (ES: 0.76 95% CI, 0.67 to 0.85) and CETAT data (0.71 95% CI, 0.60 to 0.81). AI Models that relied on EHR (0.60 95% CI, 52 to 0.69) and EMR (0.58 95% CI, 0.37 to 0.78) data showed lower sensitivity (Figure 4), which could reflect the high heterogeneity in clinical data recording, differences in definitions of ED readmission, and the quality of input variables. Although subgroup analysis showed higher pooled sensitivity for AI models using the NHIRD dataset compared to EHR- or EMR-based studies, these findings should be interpreted with caution, as only one study used NHIRD and CETAT data. EHR-based AI models were used in twenty studies, despite lower overall sensitivity and significant heterogeneity reflecting real-world variation in clinical documentation and healthcare systems. Subgroup analysis based on AI algorithms showed a significant difference (Qb = 147.35, $p < 0.001$) (Table 3), indicating the determining role of AI model selection on Sensitivity.

The highest sensitivity was for Stacking Ensemble (0.87 95% CI, 0.799 to 0.941), indicating that combining multiple algorithms can significantly increase the identification of high-risk patients. Also, C4.5, Classification Tree, DNN, and ML models had high sensitivity. The low sensitivity of BERT (0.24 95% CI, 0.087 to 0.393) was notable and could be due to the model's focus on textual data without effective integration of structured clinical data (Table 3).

FIGURE 4. FOREST PLOT SHOWED SUBGROUP META-ANALYSIS OF SENSITIVITY ACCORDING TO DATA SOURCE

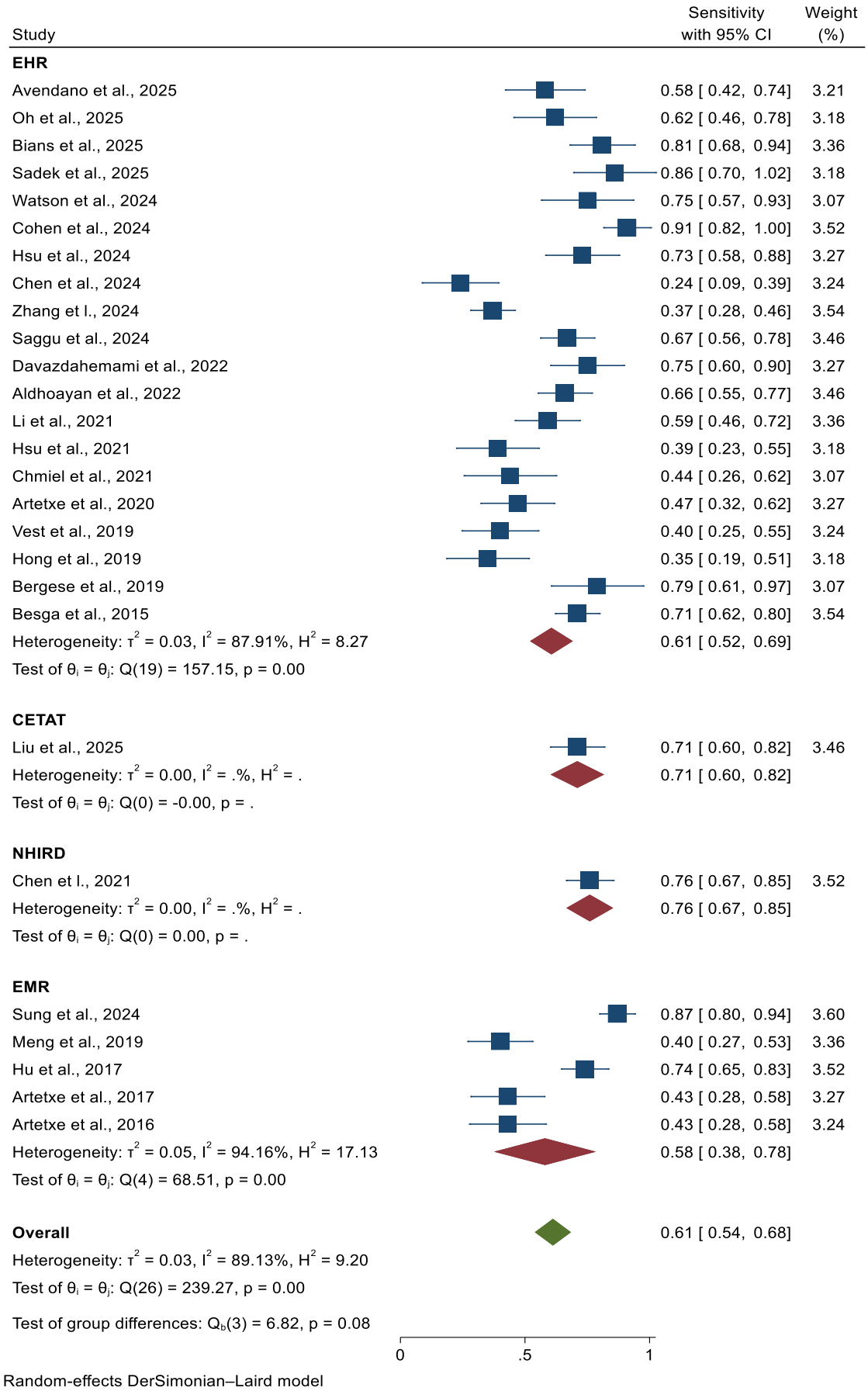
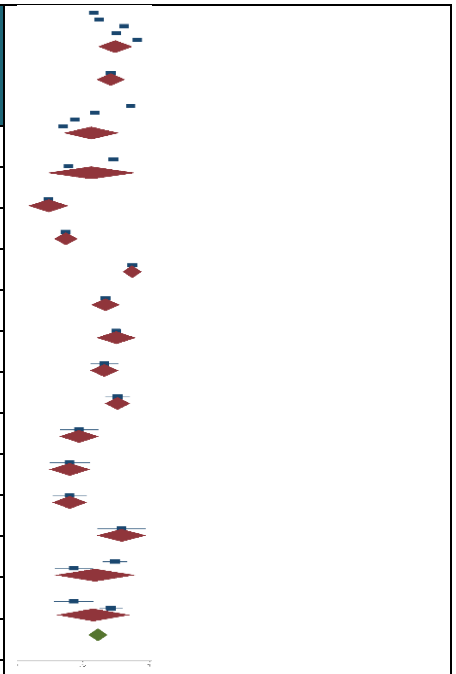


TABLE 3. SUBGROUP META-ANALYSIS OF SENSITIVITY ACCORDING TO ARTIFICIAL INTELLIGENCE ALGORITHM TYPE

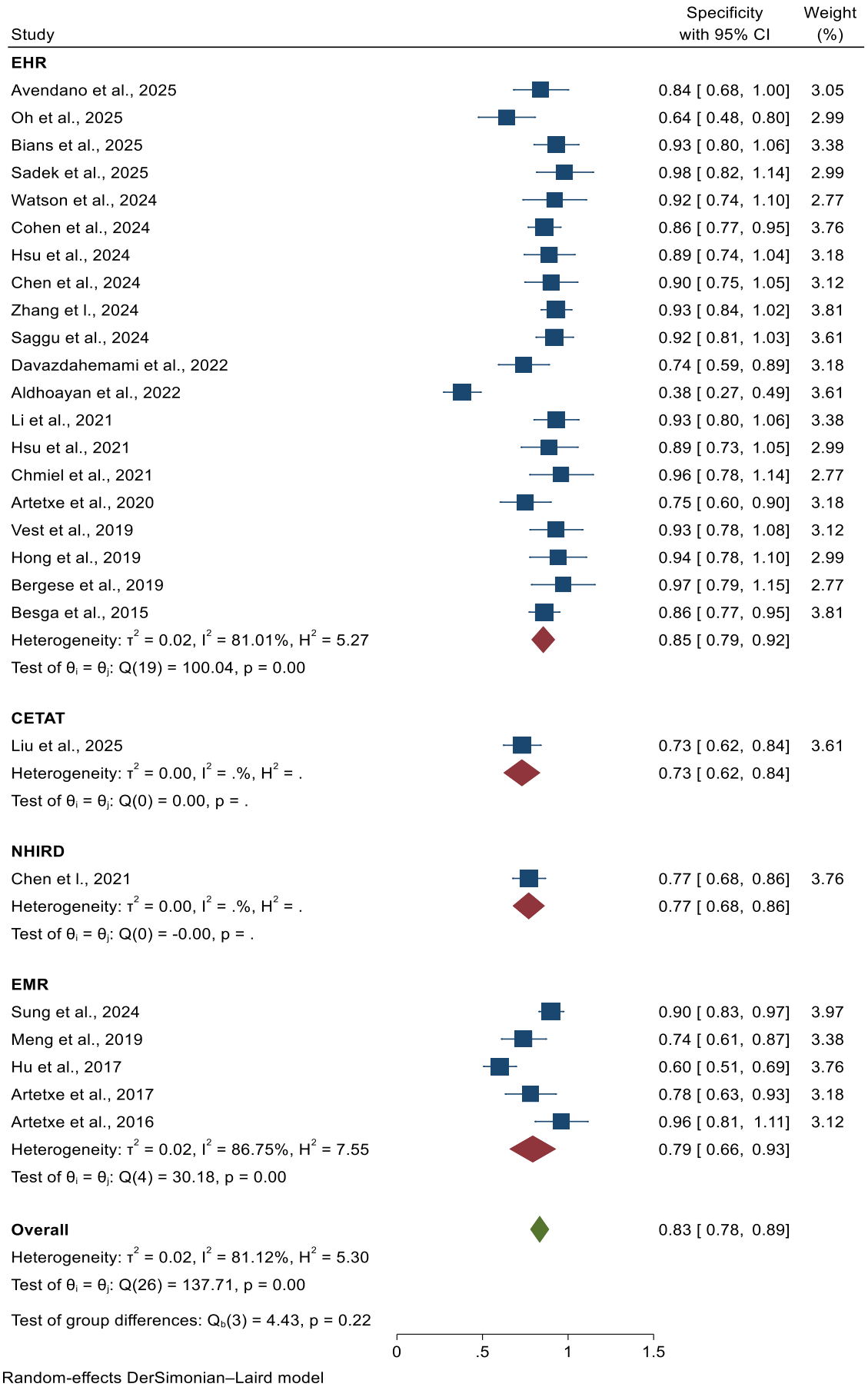
AI algorithm	Number of studies	Pooled Sensitivity	95% Confidence Interval
BERT	1	0.240	0.087 – 0.393
Bagging	1	0.470	0.323 – 0.617
C4.5	1	0.760	0.666 – 0.854
CT	1	0.790	0.606 – 0.974
CatBoost	1	0.660	0.552 – 0.768
DA	1	0.400	0.271 – 0.529
DNN	1	0.750	0.603 – 0.897
DT	1	0.400	0.247 – 0.553
GNN	1	0.670	0.562 – 0.778
LR	1	0.370	0.282 – 0.458
ML	5	0.743	0.612 – 0.874
MLP	1	0.710	0.602 – 0.818
RF	2	0.590	0.287 – 0.894
SVM	2	0.577	0.303 – 0.851
Stacking ensemble	1	0.870	0.799 – 0.941
Voting ensemble	2	0.562	0.229 – 0.895
XGBoost	4	0.562	0.352 – 0.771
Overall	27	0.612	0.539 – 0.685
Tests of group differences	df:16	Q_b= 147.35	P<0.001



In a subgroup meta-analysis by data source type, models developed using electronic health records (EHRs) showed the highest specificity (0.85; 95% CI: 0.78 to 0.92). This was followed by EMR-based studies with a pooled specificity of 0.79 (0.66–0.93) (Figure 5). The difference between subgroups in terms of data source type was not statistically significant (Q_b = 4.43, P = 0.219). These findings indicate that data source type alone did not have a decisive impact on the specificity of AI models.

Subgroup meta-analysis based on AI algorithm type showed significant differences in specificity. The difference test between subgroups in this section was statistically significant (Q_b = 106.76, P<0.001), indicating the important role of AI algorithm choice in model performance (Table 4).

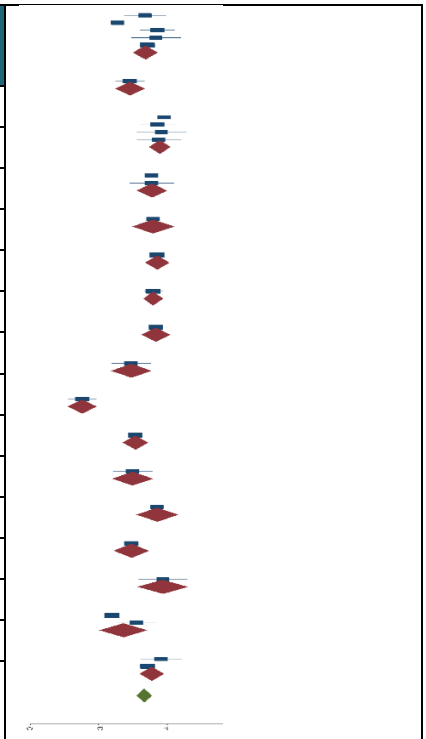
FIGURE 5. FOREST PLOT SHOWED SUBGROUP META-ANALYSIS OF SPECIFICITY ACCORDING TO DATA SOURCE



The highest specificity was related to algorithms based on gradient boosting and ensemble models, with XGBoost showing the best performance with a specificity of 0.95 (95% CI; 0.871 to1.027) and Stacking Ensemble models showing the best performance with a specificity of 0.90 (95% CI; 0.829 to 0.971). Also, some classical algorithms such as logistic regression (LR), decision tree (DT), and graph networks (GNN) also had high specificity (Table 4).

TABLE 4. SUBGROUP META-ANALYSIS OF SPECIFICITY ACCORDING TO ARTIFICIAL INTELLIGENCE ALGORITHM TYPE

AI Algorithm	No. of Studies	Pooled Specificity	95% CI
BERT	1	0.900	0.747–1.053
Bagging	1	0.750	0.603–0.897
C4.5	1	0.770	0.676–0.864
CT	1	0.970	0.786–1.154
CatBoost	1	0.380	0.272–0.488
DA	1	0.740	0.611–0.869
DNN	1	0.740	0.593–0.887
DT	1	0.930	0.777–1.083
GNN	1	0.920	0.812–1.028
LR	1	0.930	0.842–1.018
ML (unspecified)	5	0.843	0.752–0.934
MLP	1	0.730	0.622–0.838
RF	2	0.681	0.505–0.856
SVM	2	0.890	0.800–0.979
Stacking Ensemble	1	0.900	0.829–0.971
Voting	2	0.890	0.780–1.000
XGBoost	4	0.949	0.871–1.027
Tests of group differences	df:16	Q_b= 106.76	P<0.001



DISCUSSION

The findings of the present meta-analysis show that AI models have high specificity (≈0.83) in predicting patients' return to the emergency department, meaning that AI models are able to correctly identify patients who will not return to the emergency department. These results are consistent with the results of previous meta-analyses[47]. A recent meta-analysis that examined the performance of artificial intelligence algorithms in predicting emergency department readmissions showed that the average specificity in these models was about 0.92[47]. One of the notable points in the present meta-analysis was that the source of data used did not significantly affect the diagnostic specificity of AI; this finding suggests that patients who are unlikely to revisit the emergency department are generalizable across different data environments, provided that the data are standardized and reliable. However, heterogeneity between studies was high, which could be due to differences in the type of data recording, the quality of clinical variables, and the study definition of follow-up. Meta-analysis showed that specificity was consistently higher than sensitivity, with sensitivity reported at around 0.56, indicating that the models are less effective in correctly diagnosing patients with readmissions (True Positive) [48]. This observation aligns with the results of the present study; however, sensitivity was estimated to be higher in the present study. Therefore, artificial intelligence models need to improve in detecting actual patient revisits. A pooled sensitivity of 0.61 indicates that approximately 39% of patients who subsequently revisit the emergency department may not be identified by current AI models. From a clinical standpoint, such false-negative predictions may result in missed opportunities for preventive interventions, inadequate discharge planning, insufficient post-discharge follow-up, and suboptimal allocation of healthcare resources for high-risk patients. Consequently, exclusive reliance on AI-based predictions may pose potential risks to patient safety, particularly among vulnerable individuals and those with complex medical conditions.

Despite this limitation, the models included in this study demonstrated consistently high specificity, reflecting a strong capacity to accurately identify patients at low risk of emergency department revisit. Therefore, AI-based predictive models should be considered complementary clinical decision-support tools rather than replacements for physician judgment. When integrated with clinical assessment, triage findings, and professional expertise, these models may contribute meaningfully to risk stratification, optimization of resource utilization, reduction of unnecessary interventions, and enhancement of overall emergency department efficiency. Several studies have shown that using a combination of structured and unstructured data (such as clinical notes) can improve prediction performance, although its operational complexity is high and requires natural language processing skills[49]. Studies show that the use of ensemble algorithms and more complex models is preferred over simpler techniques[50, 51]. In the present study, XGBoost, Stacking Ensemble, and other hybrid approaches demonstrated superior predictive performance. This observation is consistent with previous studies suggesting that ensemble-based methods can better capture the complexity of clinical data[30, 52-56]. Many previous meta-analyses and systematic reviews on the use of AI in the emergency department have generally addressed the broad applications of AI without emphasizing quantitative measurements of the model, focusing on the transformative role of this technology in patient diagnosis and management[57]. This indicates that the field of predicting patient return to the emergency department is still relatively new and requires further studies with better reporting standards.

There were several important limitations in the present study that should be considered; high heterogeneity was observed between studies, which could be due to the AI algorithms, the criteria for defining ED revisits, and the duration of patient follow-up. Therefore, the present findings should be interpreted with caution. The number of studies in some algorithms and data sources was limited, which resulted in estimates of sensitivity and specificity with lower precision and wider confidence intervals. Therefore, more studies are needed to provide stronger evidence. Many studies used data from a single center, which limits the generalizability of the results to other populations or health systems. Also, some studies lacked full reporting of validation or multi-stage modeling, which may have led to overestimation of the true performance of the models. Finally, differences in data quality and handling of missing values between studies could also have affected the accuracy of the estimates. Furthermore, a significant geographical imbalance was observed among the included studies, with a heavy concentration of data from specific Asian healthcare databases. Future research should include more studies from Western healthcare systems would be essential to ensure the findings are generalizable to different healthcare infrastructures and patient populations.

Future research should focus on improving the identification of patients at genuine risk of emergency department revisit and utilize clinical features, textual data, and composites. To increase the generalizability of models, studies should examine multicenter data. Using structured and unstructured data simultaneously can increase predictive performance. Future research should be reported according to TRIPOD or PROBAST guidelines to allow for accurate assessment of quality and risk of bias. Based on the present results, hybrid and ensemble algorithms performed better; therefore, it is recommended to use these approaches in designing prediction models.

CONCLUSION

The present study showed that AI models demonstrated good performance in identifying patients who were unlikely to revisit the emergency department, and the overall specificity was estimated to be about 0.83. However, the sensitivity of AI models was reported to be moderate, indicating that AI models need to be improved in detecting real revisits. The type of data used did not have a significant impact on specificity and sensitivity, but the type of AI algorithm played a significant role in model performance, especially the Ensemble and XGBoost models, which provided higher performance. These findings emphasize that integrating multi-source data and using advanced hybrid algorithms can improve prediction accuracy.

References

1. Samadbeik, M., et al., Patient flow in emergency departments: a comprehensive umbrella review of solutions and challenges across the health system. BMC Health Services Research, 2024. 24(1): p. 274.

2. Berger, D., et al., Returns after discharge from the emergency department observation unit: who, what, when, and why? Western Journal of Emergency Medicine, 2023. 24(3): p. 390.
3. Salgado, R., B. Moita, and S. Lopes, Frequency and patient attributes associated with emergency department visits after discharge: retrospective cohort study. Plos one, 2022. 17(10): p. e0275215.
4. Barbic, D., et al., What is the risk of returning to the emergency department within 30 days for patients diagnosed with substance-induced psychosis? Canadian Journal of Emergency Medicine, 2022. 24(7): p. 702-709.
5. Wadhera, R.K., et al., Hospital revisits within 30 days after discharge for medical conditions targeted by the Hospital Readmissions Reduction Program in the United States: national retrospective analysis. BMJ, 2019. 366.
6. Kuttalam, S., The Role of Value-Based Care Models in Reducing Hospital Readmissions and Emergency Department Utilization: A Data-Driven Approach. Cureus. 2025. 17(5) e84761. Doi: 10-7759/cureus.844761
7. Torres, K.A., et al., Predicting the preventable: Social drivers of health on pediatric readmissions and emergency department visits. Health Psychology, 2025.
8. Pettit, R.W., et al., Artificial intelligence, machine learning, and deep learning for clinical outcome prediction. Emerging topics in life sciences, 2021. 5(6): p. 729-745.
9. Samala, A.D., S. Papadakis, and S. Rawas, Global insights into mobile learning in higher education: a PRISMA-guided bibliometric analysis from 2007 to 2023. International Journal of Educational Reform, 2025: p. 10567879251341869.
10. Tyler, S., et al., Use of artificial intelligence in triage in hospital emergency departments: a scoping review. Cureus, 2024. 16(5).
11. Jabbari, L., Dashtian, J., Azizkhani, L. CT distribution patterns of traumatic pneumothorax in blunt thoracic trauma: implications for targeted emergency imaging. Emerg Radiol, 2026.
12. Lee, J.-T., et al., Prediction of hospitalization using artificial intelligence for urgent patients in the emergency department. Scientific reports, 2021. 11(1): p. 19472.
13. Drolia, M., et al., Mobile learning applications for refugees: A systematic literature review. Education Sciences, 2022. 12(2): p. 96.
14. Zacharis, G. and S. Papadakis, Can AI Grade Like a Human? Validity, Reliability, and Fairness in University Coursework Assessment. Educational Process: International Journal, 2025. 19: p. e2025591.
15. Subasi, A., Hospital readmission forecasting using artificial intelligence, in Applications of Artificial Intelligence in Healthcare and Biomedicine. 2024, Elsevier. p. 455-520.
16. Nsama, F., Assessing the Cost-Containment Effectiveness of AI-Based Predictive Models in Reducing Avoidable Readmissions and Overtreatment in US Medicare Hospitals. Applied Sciences, Computing, and Energy, 2025. 2(2): p. 452-466.
17. Andrew, S., Revolutionizing Healthcare with Artificial Intelligence: Predictive Models for Readmission Reduction in US Hospitals. International journal of Computational Intelligence in Digital Systems, 2024. 13(01): p. 29-47.
18. Tugwell, P. and D. Tovey, PRISMA 2020. 2021, Elsevier. p. A5-A6.
19. de Jong, Y., et al., Appraising prediction research: a guide and meta-review on bias and applicability assessment using the Prediction model Risk Of Bias Assessment Tool (PROBAST). Nephrology, 2021. 26(12): p. 939-947.
20. Avendano, S.A.A., et al., Revisits, Readmission, and Mortality From Emergency Department Admissions for Older Adults With Vague Presentations: Longitudinal Observational Study. JMIR aging, 2025. 8(1): p. e55929.
21. Oh, E.G., et al., Predicting Readmission Among High-Risk Discharged Patients Using a Machine Learning Model With Nursing Data: Retrospective Study. JMIR Medical Informatics, 2025. 13(1): p. e56671.
22. Liu, Q., et al., Early identification of high-risk patients admitted to emergency departments using vital signs and machine learning. World Journal of Emergency Medicine, 2025. 16(2): p. 113.
23. Bians, N.K., et al., Predicting child and adolescent mental health emergency department revisits: a machine-learning approach compared to a clinician-derived baseline. BMC Medical Informatics and Decision Making, 2025.
24. Sadek, Z., Predicting Emergency Department Bounce Backs: A Machine Learning Approach. 2025.
25. Watson, M., et al., Performance of machine learning versus the national early warning score for predicting patient deterioration risk: a single-site study of emergency admissions. BMJ Health Care Inform, 2024. 31(1).
26. Cohen, L.J., et al., Diagnosis of the Suicide Crisis Syndrome in the Emergency Department Associated With Significant Reduction in 3-Month Readmission Rates. Journal of Clinical Psychiatry, 2024. 85(4): p. 24m15320.

27. Hsu, C.-C., et al., Machine learning models for predicting unscheduled return visits of patients with abdominal pain at emergency department and validation during COVID-19 pandemic: A retrospective cohort study. Medicine, 2024. 103(8): p. e37220.
28. Chen, T.-Y., T.-Y. Huang, and Y.-C. Chang, Using a clinical narrative-aware pre-trained language model for predicting emergency department patient disposition and unscheduled return visits. Journal of Biomedical Informatics, 2024. 155: p. 104657.
29. Zhang, Y., et al., Aspiring to clinical significance: Insights from developing and evaluating a machine learning model to predict emergency department return visit admissions. PLOS Digital Health, 2024. 3(9): p. e0000606.
30. Sung, C.-W., et al., Prediction of high-risk emergency department revisits from a machine-learning algorithm: A proof-of-concept study. BMJ health & care informatics, 2024. 31(1): p. e100859.
31. Saggu, S., et al., Prediction of emergency department revisits among child and youth mental health outpatients using deep learning techniques. BMC medical informatics and decision making, 2024. 24(1): p. 42.
32. Davazdahemami, B., P. Peng, and D. Delen, A deep learning approach for predicting early bounce-backs to the emergency departments. Healthcare Analytics, 2022. 2: p. 100018.
33. Aldhoayan, M.D. and A.M. Khayat, Leveraging advanced data analytics to predict the risk of all-cause seven-day emergency readmissions. Cureus, 2022. 14(8): p. e27630.
34. Li, J.-D., et al. Applying Problem Transformation Methods for Predicting 3-Days All-Cause Readmission in the Emergency Department. in 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI). 2021. IEEE.
35. Hsu, C.-C., et al., A machine learning model for predicting unscheduled 72 h return visits to the emergency department by patients with abdominal pain. Diagnostics, 2021. 12(1): p. 82.
36. Chmiel, F., et al., Using explainable machine learning to identify patients at risk of reattendance at discharge from emergency departments. Scientific reports, 2021. 11(1): p. 21513.
37. Chen, R.-F., et al., Predicting unscheduled emergency department return visits among older adults: population-based retrospective study. JMIR Medical Informatics, 2021. 9(7): p. e22491.
38. Artetxe, A., et al., Balanced training of a hybrid ensemble method for imbalanced datasets: a case of emergency department readmission prediction. Neural Computing and Applications, 2020. 32(10): p. 5735-5744.
39. Vest, J.R. and O. Ben-Assuli, Prediction of emergency department revisits using area-level social determinants of health measures and health information exchange information. International journal of medical informatics, 2019. 129: p. 205-210.
40. Meng, F., et al., Predicting 72-hour reattendance in emergency departments using discriminant analysis via mixed integer programming with electronic medical records. Journal of Industrial & Management Optimization, 2019. 15(2).
41. Hong, W.S., A.D. Haimovich, and R.A. Taylor, Predicting 72-hour and 9-day return to the emergency department using machine learning. JAMIA open, 2019. 2(3): p. 346-352.
42. Bergese, I., et al., An innovative model to predict pediatric emergency department return visits. Pediatric emergency care, 2019. 35(3): p. 231-236.
43. Hu, Y.-H., et al., Predicting return visits to the emergency department for pediatric patients: applying supervised learning techniques to the Taiwan National Health Insurance Research Database. Computer methods and programs in biomedicine, 2017. 144: p. 105-112.
44. Artetxe, A., et al. Emergency department readmission risk prediction: A case study in Chile. in International Work-Conference on the Interplay Between Natural and Artificial Computation. 2017. Springer.
45. Artetxe, A., et al. Predicting 30-day emergency readmission risk. in International Workshop on Soft Computing Models in Industrial and Environmental Applications. 2016. Springer.
46. Besga, A., et al., Risk factors for emergency department short time readmission in stratified population. BioMed research international, 2015. 2015(1): p. 685067.
47. Kuo, K.-M. and C.S. Chang, A meta-analysis of the diagnostic test accuracy of artificial intelligence predicting emergency department dispositions. BMC Medical Informatics and Decision Making, 2025. 25(1): p. 187.
48. Kuo, K.-M., W.-S. Wu, and C.S. Chang, A Meta-Analysis of the Diagnostic Test Accuracy of Artificial Intelligence for Predicting Emergency Department Revisits. Journal of Medical Systems, 2025. 49(1): p. 81.
49. Juang, W.-C., et al., Assessing Revisit Risk in Emergency Department Patients: Machine Learning Approach. JMIR AI, 2025. 4(1): p. e74053.

50. Zhou, Z.-H., Ensemble methods: foundations and algorithms. 2025: CRC press.
51. Jabbari, L., Azizkhani, L., Rouzbin, F., Salarpour, E., Hayati, S. Is admission neutrophil-to-lymphocyte ratio an outcome predictor in critically ill multiple trauma patients? A retrospective single-centre cross-sectional study. Trauma, 2025, 27(2): p.128-133.
52. Berikol, G.B., et al., Mapping artificial intelligence models in emergency medicine: A scoping review on artificial intelligence performance in emergency care and education. Turkish Journal of Emergency Medicine, 2025. 25(2): p. 67-91.
53. Brasen, C.L., et al., Machine learning in diagnostic support in medical emergency departments. Scientific Reports, 2024. 14(1): p. 17889.
54. Olawade, D.B., et al., Smart emergency care: a narrative review of predictive machine learning models. Annals of translational medicine, 2025. 13(5): p. 60.
55. Alhazmi, H.N., A Prediction Triage System for Emergency Department During Hajj Period using Machine Learning Models. International Journal of Computer Science & Network Security, 2024: p. 11-23.
56. Brasen, C.L., et al., Machine learning in diagnostic support in medical emergency.
57. Alimiri Dehbaghi, H. and K. Khoshgard, Revolutionizing emergency care: an overview of the transformative role of artificial intelligence in diagnosis, triage, and patient management. International Journal of Emergency Medicine, 2025. 18(1): p. 242.