MAYO CLINIC PROCEEDINGS: DIGITAL HEALTH

REVIEW

Predicting Inpatient Admissions From Emergency Department Triage Using Machine Learning: A Systematic Review

Ethan L. Williams, MD; Daniel Huynh, MD; Mohamed Estai, MBBS, PhD; Toshi Sinha, PhD; Matthew Summerscales, MBBS; and Yogesan Kanagasingam, PhD

Abstract

This study aimed to evaluate the quality of evidence for using machine learning models to predict inpatient admissions from emergency department triage data, ultimately aiming to improve patient flow management. A comprehensive literature search was conducted according to the PRISMA guidelines across 5 databases, PubMed, Embase, Web of Science, Scopus, and CINAHL, on August 1, 2024, for English-language studies published between August 1, 2014, and August 1, 2024. This yielded 700 articles, of which 66 were screened in full, and 31 met the inclusion and exclusion criteria. Model quality was assessed using the PROBAST appraisal tool and a modified TRIPOD+AI framework, alongside reported model performance metrics. Seven studies demonstrated rigorous methodology and promising in silico performance, with an area under the receiver operating characteristic ranging from 0.81 to 0.93. However, further performance analysis was limited by heterogeneity in model development and an unclear-to-high risk of bias and applicability concerns in the remaining 24 articles, as evaluated by the PROBAST tool. The current literature demonstrates a good degree of in silico accuracy in predicting inpatient admission from triage data alone. Future research should emphasize transparent model development and reporting, temporal validation, concept drift analysis, exploration of emerging artificial intelligence techniques, and analysis of real-world patient flow metrics to comprehensively assess the usefulness of these models.

 ■ Mayo Clin Proc Digital Health 2025;3(1):100197

From the School of Medicine, The University of Notre Dame, Fremantle, Western Australia, Australia (E.L.W., T.S., Y.K.); Emergency Department, St John of God Midland Public and Private Hospitals, Midland, Western Australia, Australia (E.L.W., M.S., Y.K.); General Medicine Department, Royal North Shore Hospital, St Leonards, New South Wales, Australia (D.H.); and School of Human Sciences, The University of Western Australia, Crawley, Western Australia, Australia (M.E.).

Hospital Overcrowding

The emergency department (ED) is a crucial hospital department, diagnosing, treating, and determining the trajectory of a patient's health care journey. The complexity and number of presentations to ED are increasing globally, even relative to population growth, putting additional pressure on health care systems.[1,2] This leads to overcrowding, inefficiency, and worse patient outcomes.[1,2] Hospital factors contributing to overcrowding can be split into 3 categories: input, throughput, and output, with each stage impacting patient flow and contributing to overcrowding.[3] Currently, patient flow is often the role of a patient flow coordinator, usually a senior nurse or a group of nurses. They try to synthesize the large quantity of ever-changing data to optimize patient flow through a hospital, with increasing research aiming to innovate this process.[4] Given the progressive digitalization of health care systems, there is an increasing ability to provide additional tools to patient flow coordinators to improve the prediction of patient throughput and output.

Machine Learning Development Overview

Commonly used models discussed in this field and review include logistic regression, decision trees, random forests, gradient-boosted machines, extreme gradient boosting (XGB), and neural networks (NNs). A visual representation of these models is displayed in Figure 1, and a background explanation of each model is outlined in Supplemental Appendix A (available online at https://www.mcpdigitalhealth.org/).

The development of these models, once data have been collected, prepared, and cleaned, shares a common general structure: training, internal validation, and testing. This

ARTICLE HIGHLIGHTS

- Admission prediction models have demonstrated an AUROC of 0.81-0.93 during ‘in-silico’ testing, though have not yet been validated in a live clinical setting.
- Routinely collected features at triage such as age, triage category, arrival mode and vital signs often contribute most to predictive performance.
- XGB and MLP models are the most common high performing models in the currently published literature- Beyond model structure and feature selection, population and healthcare system characteristics appear to greatly impact model performance.
- Future research should explore temporal validation and integration of variety of model designs, such as transformer networks.

process is critical to reduce model bias and variance. Bias refers to the error that occurs when a model is too simplistic and fails to capture the underlying patterns in the data. While variance refers to the model’s sensitivity to small fluctuations in the training data, making it too specific to the training set. Generally, high bias leads to underfitting and high variance to overfitting, both of which are undesirable. This can be visualized in Figure 2.

In the training phase, a large subset of the data, often selected randomly, is used to train the model based on the provided features, also known as predictors or independent variables. The model aims to optimize its performance by adjusting its parameters to minimize prediction error. Developers may also perform additional steps, such as feature selection, handling missing data, modifying the model structure, or addressing class imbalance. However, some models, such as XGB, can handle missing data and class imbalance natively, reducing the need for additional techniques like undersampling or oversampling.

The internal validation phase involves exposing the model to a mutually exclusive validation set (held-out set) to tune its hyperparameters. Hyperparameters are the parameters that control the model’s structure and behavior. This phase is essential to prevent overfitting. Alternate techniques such as k-fold crossvalidation are increasingly used to ensure robustness. During this phase of model development, the validation set is used for internal validation of the model. This importantly differs from the validation phase of model implementation, used commonly in literature to reference external validation of the model. For example, the PROBAST tool refers to a validation study as a study that performs external validation of the fully developed model, once it has already completed the 3 phases of development that we outline in this review.[5] The difference between internal validation and external validation is important to consider when analyzing the literature and model development because the single-term validation is often used referring to both internal and external validation.

The final testing stage is when the model is tested on another held-out set, where the performance can be accurately reported with no further adjustment to hyperparameters. This aims to provide an unbiased estimate of the model’s generalizability to new data. There are many variations on how to perform each phase, especially training and validation, with each phase being important to reduce bias and variance and predict model performance capabilities.

Machine Learning in the ED

Machine learning (ML) and artificial intelligence (AI) in the ED are a growing field, spanning patient predictions, diagnosis, clinical decision making, and operational efficiency.[6,7] The theoretical benefit of improved hospital efficiency, improved patient flow, and improved patient outcomes is widely discussed but requires further evidence.[8] Current barriers to widespread implementation include variable quality of input data, proprietary models lacking transparent validation, medicolegal regulations, and the imperfect probabilistic nature of model decisions.[7,8] Therefore, although various ML models have attempted to predict hospital admission from triage data, many lack clinical implementation.[9,10] Bridging the gap between theoretical in silico modeling and implementation, while addressing the barriers faced by ML models, is critical to assess whether this technology can improve patient care outcomes.

Mayo Clin Proc Digital Health ■ March 2025;3(1):100197 ■ https://doi.org/10.1016/j.mcpdig.2025.100197
www.mcpdigitalhealth.org

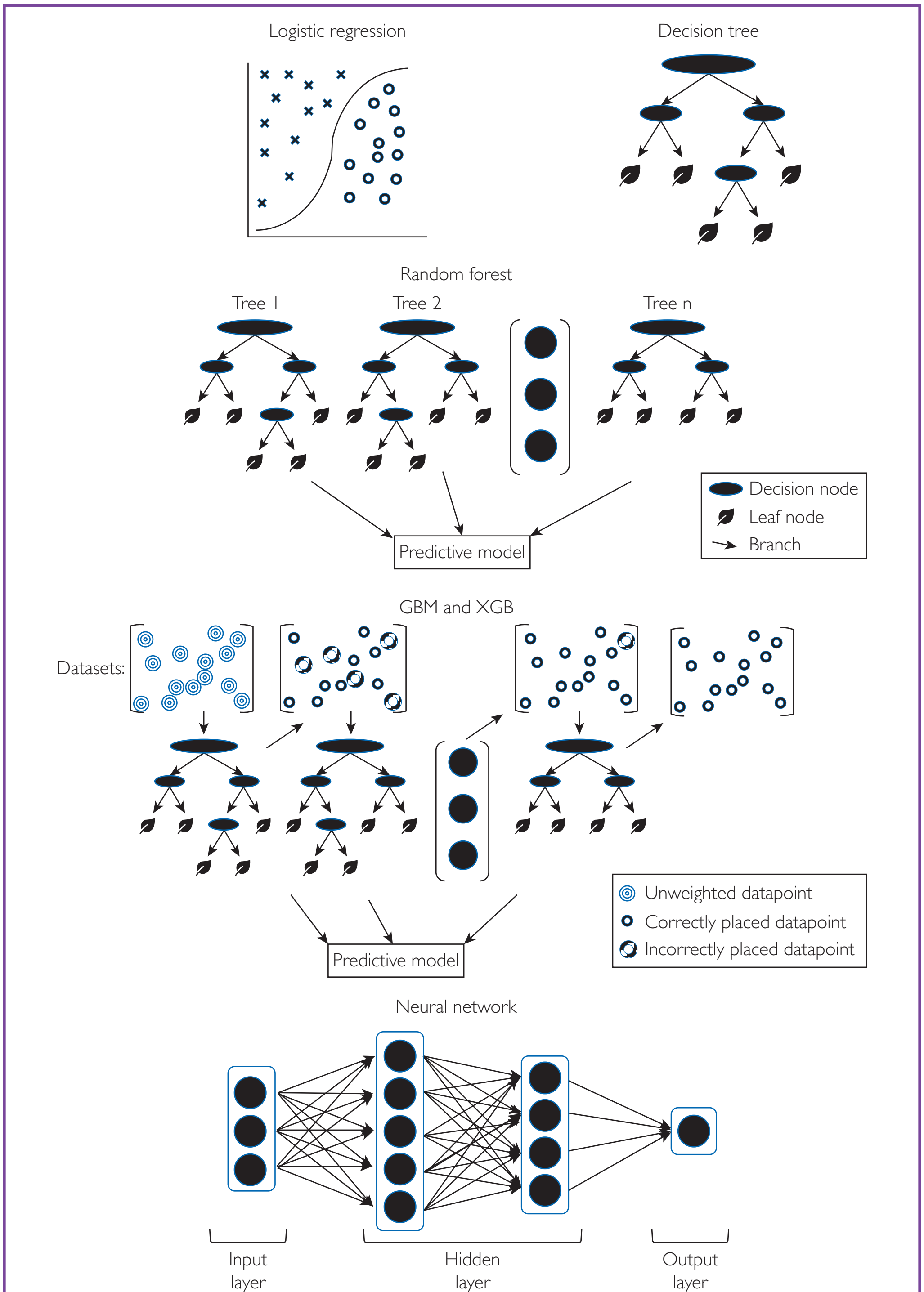

FIGURE 1. Visualization of common machine learning techniques used for admission prediction. GBM, gradient boosted machine; XGB, extreme gradient boosting.

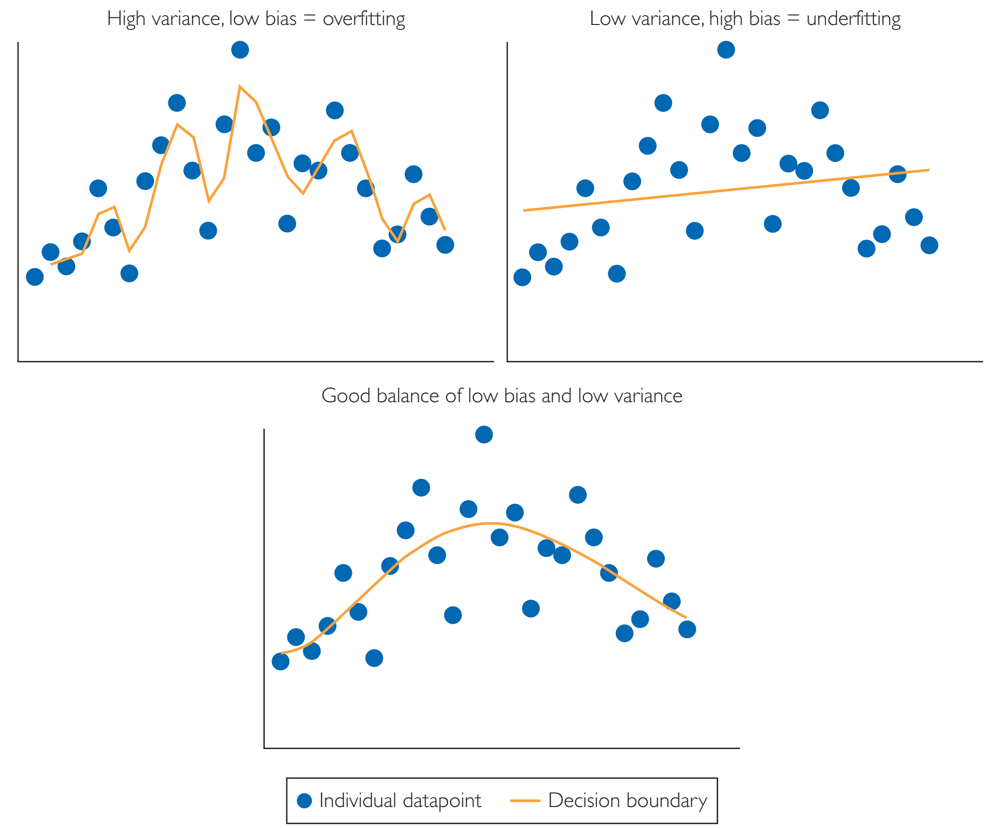

FIGURE 2. Visualization of model variance and bias.

Aim

To our knowledge, there is no recent review assessing the quality of evidence for ML predicting inpatient admission from an ED setting, nor a review assessing the currently limitations or areas for further development required for implementation. This review aimed to analyze the current literature to map the landscape of ML and AI in ED admission prediction from data available at triage. Analysis was performed using the critical appraisal of each model's development, with presentation of each model's performance in context of its appraisal. Secondarily, we aimed to determine the strengths, weaknesses, and gaps required to address in admission prediction modeling to optimize models for real-world implementation to improve patient flow, hospital efficiency, and subsequently patient outcomes.

METHODS

Study Design

This study was designed and completed by a research team comprising 4 clinicians (E.L.W., D.H., M.S., and M.E.), a software engineer with a PhD in ML (T.S.), and a PhD supervisor with previous ML experience (Y.K.). The literature search was broad, with comprehensive screening and appraisal performed independently by clinicians E.L.W. and D.H. before it was shared with the wider research team, as described further. The inclusion and exclusion criteria were first applied to the abstract and then the full text, before appraisal with the PROBAST tool and a modified TRIPOD+AI tool was completed for each article's best performing eligible model.[5,11,12] Model performance was determined by reported area under the receiver operator characteristic curve (AUROC). When unavailable, accuracy as a percentage was reported. Study characteristics and performance measures were extracted manually during completion of the appraisal tools. Results were presented as a narrative synthesis according to search results, study characteristics, the 4 PROBAST domains, and the ethical TRIPOD+AI domain.[5,11,12] Model performance measures are presented alongside the PROBAST critical appraisal rating in the discussion to guide the relative confidence in the reported performance. Owing to significant heterogenicity in model development and analysis, a meta-analysis was not performed. Heterogenicity was handled by only comparing methodology and performance in the discussion for articles with low risk of bias and low applicability concerns, to improve the robustness of reported findings. This study and the methodology were registered before analysis with PROSPERO (CRD42024580183).

Inclusion Criteria

The inclusion criteria were as follows:

- English full-text article available.
- Peer-reviewed journal article.
- Articles predicting the inpatient admission of patients from information available at triage. Inpatient admission is defined as admission to any inpatient ward, including emergency short stay units, from the ED. Admissions to inpatient wards from routes other than triage are not included.
- Articles including all presentations to either an adult or mixed adult-pediatric ED. Articles excluding pediatric patients in a mixed ED were still included.
- Articles published within the past 10 years from the search date.

Exclusion Criteria

The exclusion criteria were as follows:

- Lack of admission prediction based on patient data from triage.
- Review and conference articles.
- Articles focusing on only 1 disease, presenting complaint or patient subgroup (including exclusion of pediatric-only articles).
- Articles that only present the performance of a model trained on data collected after triage.
- Articles where the outcome is not hospital admission.

Search Strategy

A literature search was performed by an experienced researcher (M.E.) on 5 databases: PubMed, Embase, Web of Science, Scopus, and CINAHL on August 1, 2024. The search criteria were as follows: "*machine learning*" OR "*artificial intelligence*" OR "*deep learning*" OR "*neural network**" OR "*support vector machine**" OR "*random forest**" OR "*gradient boosting*" OR "*gradient boosted*" OR "*decision tree**" OR "*logistic regression*" AND "*emergency department*" OR "*ED*" OR "*emergency ward*" OR "*emergency room*" OR "*ER*" OR "*accident and emergency*" OR "*triage*" AND "*admission prediction*" OR "*admission forecasting*" OR "*hospital admission*." The search, performed by an author (M.E.), yielded 700 unique search results over the past 10 years, with the results shared with the entire research team.

Article Screening

Abstract review was completed by the lead author (E.L.W.) independently, with eligibility defined as any article that did not meet any exclusion criteria and possibly met the inclusion criteria. E.L.W. then analyzed the full-text independently and sent all selection decisions to the second author (D.H.) for crossvalidation and verification of selection, before sharing screening results with the wider research team. All remaining journal articles were included in critical appraisal. This selection process is outlined in Figure 3, according to the PRISMA guidelines.[13]

Critical Appraisal and Methodology

The PROBAST tool was selected as the primary critical appraisal for its rigor in assessing the risk of bias and applicability of diagnostic and prognostic prediction models.[5,12] Four questions from the TRIPOD+AI critical appraisal tool were added to the PROBAST tool to increase relevance of the appraisal tool for models aiming for real-world hospital implementation.[11] The lead and second author independently completed the appraisal tool for each of the 31 articles. After summating notes together on the appraisal tool, any points of difference between these 2 reviewers

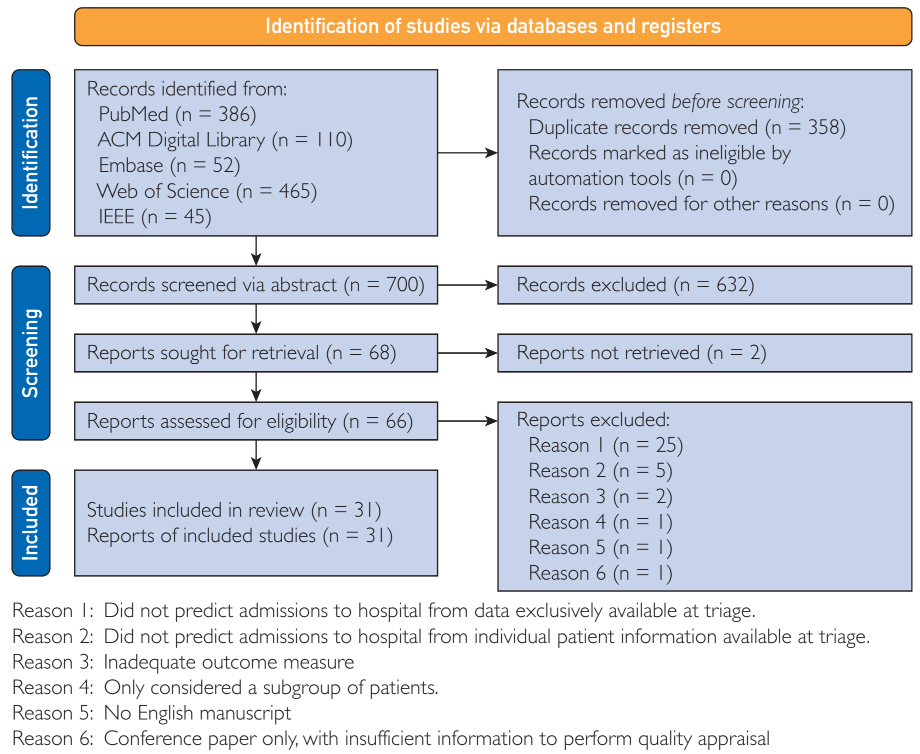

FIGURE 3. Literature search flow diagram. Adapted from the PRISMA template.[13]

were raised with the extended research team for discussion and, if required, a consensus vote. The common themes were collaboratively extracted by E.L.W. and D.H. and then shared with the wider research team throughout the narrative synthesis, which was completed by E.L.W. The entire appraisal document is available as Supplemental Appendix B (available online at https://www.mcpdigitalhealth.org/).

RESULTS

Search Results

The search resulted in 700 articles published between August 1, 2014, and August 1, 2024, on the PubMed, Embase, Web of Science, Scopus, and CINAHL databases. After abstract review, 68 articles remained, with 66 full-text articles successfully retrieved and 31 articles passing full-text review.[14–44] The search results are reported according to the PRISMA guidelines in Figure 3. The full list of excluded articles, including reason for exclusion, is reported in Supplemental Appendix C (available online at https://www.mcpdigitalhealth.org/).

Study Characteristics

Thirty-one studies were included in this review, with the best performing model that met inclusion criteria being analyzed. Of these studies' best performing models, 11 included XGB models, 10 NN models, 6 natural language processing (NLP), and 5 gradient-boosted machine models. Fourteen studies used patient data from the United States, with 3 from Australia and 2 each from England, France, Spain and the Republic of Korea. The number of presentations, period of data collection, and number of centres were highly variable between studies (Table 1). Admission rates varied from 11% to 45%, with an average admission rate of approximately 24·6%, unadjusted for sample sizes. Average admission rates in the United States and Spain were lower at 19.3% and 12.3%

TABLE 1. Study Demographics[a]

Reference, year	No. of presentations	Country of origin	Percentage admitted	Period of presentations	No. of centres	Model type	Reported ethics approval	Code available	Data set available
Ahmed et al,[14] 2022	453,664	USA	∼25[b]	2017-2019	4	Tabu-Adaboost	No	No	No
Akhlaghi et al,[15] 2024	77,125	Australia	23.9	2021-2022	1	NLP	Yes	No	Yes
Araz et al,[16] 2019	118,005	USA	24.7[b]	2006-2009	1	XGB	No	?[c]	No
Chen et al,[17] 2022	828,689	Taiwan	18.3	2014-2019	1	NLP + AB-NN	No	No	No
Cusido et al,[18] 2022	3,189,204	Spain	11.0	2018	Regional data set	GBM	No	No	No
De Hond et al,[19] 2021	158,799	Netherlands	40	2017-2019	National data set	XGB	Yes	Yes	No
Duckworth et al,[20] 2021	82,402	England	35-45[b]	2019-2020	1	XGB	Yes	No	Yes
Fakhfakh-Maala et al,[21] 2022	225,392	France	41	2016-2020	1	MLP	Waived	No	No
Feretzakis et al,[22] 2024	276,939	Israel	Not reported	2011-2019	1	GBM	No	?[c]	No
Glicksberg et al,[23] 2024	864,089	USA	18.5	Not reported	7	GPT-4 + XGB	No	?[c]	No
Graham et al,[24] 2018	107,545	Northern Ireland	24.7	2015	2	GBM	No	?[c]	No
Handley et al,[25] 2015	159,200	USA	28.4	2007-2010	1	MLP	Yes	No	No
Hong et al,[26] 2018	560,486	USA	29.7	2013-2017	3	XGB	Yes	Yes	Yes
Huo et al,[27] 2019	100,000	USA	Not reported	2014-2017	3	XGB + ET	No	Yes	Yes
Kim et al,[28] 2022	27,747	Republic of Korea	21.8	2018-2019	1	XGB	Yes	No	No
King et al,[29] 2022	213,969	England	16-18	2019-2021	1	XGB	Waived	Yes	Yes
Kishore et al,[30] 2023	847,983	Australia	41.6	2015-2022	1	XGB	Yes	?[c]	Yes
Mehra et al,[31] 2024	52,222	Switzerland	Not reported	2019-2021	1	NLP	Yes	No	No
Monahan et al,[32] 2022	93,847	USA[b]	13.5	2019	1	LR	Yes	No	No
Park et al,[33] 2023	154,383	Republic of Korea	25.2	2016-2020	1	MLP	Yes	?[c]	No
Raita et al,[34] 2019	135,470	USA	16.2	2007-2015	National database	MLP/GBM	Yes	?[c]	Yes
Rendell et al,[35] 2019	1,721,294	Australia	40.7	2013-2014	Regional database	k-NN	Yes	No	No
Ryu et al,[36] 2022	144,827	USA	18-49	2018-2019	6	XGB	Yes	No	No
Sax et al,[37] 2023	5,315,176	USA	12.7	2016-2020	21	MLP	No	No	No
Somanchi et al,[38] 2022	213,501	USA[b]	17	67 mo, unknown dates	1	GBM	No	No	No

Continued on next page

TABLE 1. Continued

Reference, year	No. of presentations	Country of origin	Percentage admitted	Period of presentations	No. of centres	Model type	Reported ethics approval	Code available	Data set available
Sterling et al,[39] 2019	256,878	USA	26.5	2015-2016	3	NLP	Yes	No	No
Tschoellitsch et al,[40] 2023	1506	Austria	37.9	10 d, 2019-2020	1	MLP/XGB[b]	Yes	?[c]	Yes
Uhl et al,[41] 2022	84,865	France	38.9	2020-2021	2	MLP	No	?[c]	No
Yao et al,[42] 2020	118,602	USA	11.6	2012-2016	National database	CNN	No	No	Yes
Zhang et al,[43] 2017	47,200	USA	13.4	2012-2013	National database	LR + NLP	No	?[c]	Yes
Zlotnik et al,[44] 2016	255,668	Spain	13.5	2011-2012	1	MLP	Yes	No	No

[b]Not explicitly stated.

[c]Transparent model development but hyperparameters not available.

aAB-NN, attention based neural network; CNN, convolutional neural network; ET, embedding techniques; GBM, gradient boosted machine; k-NN, K nearest neighbor; LR, logistic regression.

respectively, and higher in France and Australia at 40.0% and 35.4%, respectively.

Participants

Critical appraisal of participant inclusion had a low risk of bias for 27 articles included in this review. Three articles did not report any explicit inclusion or exclusion criteria for patients included which created an unclear risk of bias.[21,27,36] A single article only drew a small number of participants from a brief period of 10 days, with exclusion of critical patients, reducing real-world applicability.[40] While greater demographic details would aid in contextualizing and analyzing articles, participant data collection and inclusion were otherwise done appropriately according to the PROBAST tool.

Predictors

Predictors of admission, often referred to as features, had a low concern for bias and good applicability in 27 studies. For the studies with a unclear or high level of concern, 2 did not share their full list of features[21,42]: 1 included a medical history, which was unclear whether taken at triage or later in the patient visit,[21] and 1 article included patient waiting time,[22] a metric likely not available on presentation. Applicability concern was also high for the study by Tschoellitsch et al[40] because they used a single medical student observer, with subjective observational measures, rather than an experienced triage nurse as per standard ED practice. For the remaining articles, features were timely and consistently defined for all participants within the constraints of a real-world ED. These features were similar between articles, including demographics, arrival time and mode, patient vital signs, and occasionally triage notes or a medical history at the center. No models including blood results, imaging, in-hospital echocardiogram results, or physician medical notes as a feature in their model were included in this review given the exclusion criteria.

Outcome

The outcome measure of accuracy of the prediction of admission to hospital via an ED, reported as AUROC, was the key outcome of interest. Although admission prediction was also the key outcome for most studies included, 7 did not give a clear outcome definition,[18,22,25,27,36,38,42] and many more did not discuss handling of alternative disposition. Alternate dispositions include patients who were transferred, "did not wait," or passed away in the ED. This is important as some studies would consider transfers as an admission, because they met medical criteria for admission but required transfer for other reasons. However, for models drawing training data from multiple hospitals, this can result in double counting of transferred patients if the model is trained or tested on both hospitals the patient visited. Some studies included death in ED as an admission, others excluded the patient from the data set. Many studies excluded any data for patients who did not wait or left against medical advice. Although this definition could introduce bias, there is no current gold standard for handling of these dispositions. Therefore, unclear or high bias ratings were only given to articles lacking a clear outcome definition or due to other clear sources of bias.

A key issue identified in the outcome applicability domain pertains to the training and testing sets. Twenty-one studies used outcomes for the training and testing data sets that were derived from within the same period, often via randomization.[16–19,21,22,24,26–28,31–37,39,40,42,43] Although this was good for initial model development, if only outcomes from the same period were included, no prospective predictive ability was demonstrated. Given this being the primary purpose in real-world clinical application, articles with no prospective predictive ability were classed as unclear applicability. Overall, 8 studies of the 31 had appropriate outcome measure definition and an appropriate temporal split between the training and testing sets.[14,15,20,23,29,30,41,44]

Analysis

In the domain of data analysis, 12 models were scored as unclear or high risk of bias. Although some ML model types such as XGB can handle cases with missing data natively, 2 articles with models requiring processing of cases with missing data did not report handing of any missing data.[17,27] Moreover, 3 articles excluded any patients with missing data, reducing real-world applicability.[18,32,35] Lack of a discussed internal validation technique or validation set also introduced bias in 4 articles.[22,23,31,42] Conversely, 5 articles extracted their results from internal validation results, rather than a held-out test set, introducing bias.[27,32,35,40,43]

A summary of these articles' best performing model results is displayed in Table 2 alongside the overall appraisal tool ratings. The fully completed appraisal tool for each article can be found in Supplemental Appendix B. Model performance and useability in context of the appraisal are discussed in further detail in the Discussion section.

Ethics and Data Availability

Successful ethics registration was reported by 16 articles, whereas 2 reported ethics being waived for their study and the remaining 13 not reporting any ethics application or approval. Four articles made their code publicly available; 10 articles clearly explained model development although did not share hyperparameters, and the 17 remaining articles had a limited degree of code sharing. Patient data are either publicly available or available on request for 10 articles, whereas the remaining 21 articles had no comments regarding data availability. Table 1 presents the ethics, code, and data availability for each article.

Feature Importance

Many features are used commonly in these models, with demographics, triage scores, and arrival mode being the most frequent used feature sets. Fewer studies also included time of presentation, vital signs, triage free text, medical history, and previous number of presentations. From studies reporting feature importance, age, lower triage score, arrival by ambulance, abnormal vital signs, and shortness of breath had the greatest positive impact on model performance in descending order. A list of included feature sets for each article can be found in Table 3.

DISCUSSION

The aim of this review was to appraise the quality of the literature demonstrating admission prediction with ML from patient data available at triage in an ED setting. The secondary objective was to explore the performance and usability of each model, within the context of their robustness as determined via the PROBAST and TRIPOD+AI assessment. Admission prediction is not a new phenomenon created by the evolution of advanced ML algorithms; clinical settings often rely on staff to predict patients' admission risk on top of their existing workload. Trotzky et al[45] demonstrated ED triage nurses

TABLE 2. PROBAST Critical Appraisal and Model Performance

Reference, year	Bias				Applicability			Overall		AUROC	
	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	Bias	Applicability	Int	Temp
Ahmed et al,[14] 2022	X	X	X	-	X	X	X	-	X	0.954	
Akhlaghi et al,[15] 2024	X	X	X	X	X	X	X	X	X		74%[a]
Araz et al,[16] 2019	X	X	X	X	X	X	?	X	?	0.863	
Chen et al,[17] 2022	X	X	X	?	X	X	?	?	?	96%	
Cusido et al,[18] 2022	X	X	?	?	X	X	?	?	?	0.894	
De Hond et al,[19] 2021	X	X	X	X	X	X	?	X	?	0.84	
Duckworth et al,[20] 2021	X	X	X	X	X	X	X	X	X	0.856	0.826
Fakhfakh-Maala et al,[21] 2022	?	?	X	-	X	-	?	-	-	0.89	
Feretzakis et al,[22] 2024	X	-	-	-	X	-	?	-	-	0.715	
Glicksberg et al,[23] 2024	X	X	X	X	X	X	X	X	X		0.88
Graham et al,[24] 2018	X	X	X	X	X	X	?	X	?	0.859	
Handley et al,[25] 2015	X	X	?	X	X	X	X	?	X	0.860	
Hong et al,[26] 2018	X	X	X	X	X	X	?	X	?	0.874	
Huo et al,[27] 2019	?	X	?	-	X	X	?	-	?	0.922	
Kim et al,[28] 2022	X	X	X	X	X	X	?	X	?	0.933	
King et al,[29] 2022	X	X	X	X	X	X	X	X	X	0.82	0.81
Kishore et al,[30] 2023	X	X	X	X	X	X	X	X	X		0.93
Mehra et al,[31] 2024	X	X	X	-	X	X	?	-	?	0.84	
Monahan et al,[32] 2022	X	X	X	?	X	X	?	?	?	0.841	
Park et al,[33] 2023	X	X	X	X	X	X	?	X	?	0.75	
Raita et al,[34] 2019	X	X	X	X	X	X	?	X	?	0.82	
Rendell et al,[35] 2019	X	X	X	?	X	X	?	?	?	0.827	
Ryu et al,[36] 2022	?	X	?	X	X	X	?	?	?	0.88	
Sax et al,[37] 2023	X	X	X	X	X	X	?	X	?	0.87	
Somanchi et al,[38] 2022	X	X	?	X	X	X	X	?	X	0.868	
Sterling et al,[39] 2019	X	X	X	X	X	X	?	X	?	0.785	
Tschoellitsch et al,[40] 2023	X	X	X	-	?	?	?	-	?	0.80	
Uhl et al,[41] 2022	X	X	X	X	X	X	X	X	X	0.86	0.87
Yao et al,[42] 2020	X	?	?	?	X	?	?	?	?	0.860	
Zhang et al,[43] 2017	X	X	X	?	X	X	?	?	?	0.846	
Zlotnik et al,[44] 2016	X	X	X	X	X	X	X	X	X	0.863	0.858

X, low concern; ?, unclear degree of concern; -, high concern; Int, internal validation results; Temp, temporal validation results.
[a]AUROC not reported in manuscript so accuracy performance reported instead.

could predict admission rates with 77% accuracy; however, accuracy significantly varied between patient cohorts. For example, acute, nonambulatory patients, triaged as level 3 according to the Canadian Triage and Acuity Scale, were more difficult to predict with an accuracy of 52%.[45] Overall, the accuracy of triage nurse admission prediction from triage is reportedly between 73% to 80.6%.[45-47] Although most articles do not report accuracy, rather AUROC, it is worth noting a baseline of 73%-80% accuracy from triage nurses may be a suggested minimum benchmark for a model to be deemed effective.

Model Performance

Performance varied significantly between the models examined in this review. AUROC ranged from 0.715 to 0.954, although when considering only articles with low risk of bias and low concern regarding applicability, hereby referred to as set A,[15,20,23,29,30,41,44] the

TABLE 3. Included Features for Each Study

Reference, year	Demographics	Arrival time	Arrival mode	Presenting complaint	Triage score	Vital signs	Triage free-text	Past presentations or admissions	Medical history	Most important features	AUROC
Ahmed et al,[14] 2022	X	X	-	X	-	X	-	-	-	Saturations, age, RR	0.954
Akhlaghi et al,[15] 2024	X	-	X	-	X	-	NLP	-	-	-	74%[a]
Araz et al,[16] 2019	X	X	X	X	X	-	-	-	-	Age, TS, arrival mode	0.863
Chen et al,[17] 2022	X	X	X	X	X	X	NLP	-	X	-	0.94
Cusido et al,[18] 2022	X	-	-	X	X	-	-	X	-	Triage complaint, TS, age	0.894
De Hond et al,[19] 2021	X	X	X	X	X	X	-	-	-	-	0.84
Duckworth et al,[20] 2021	X	X	X	X	-	X	-	-	X	PMH, arrival mode, age	0.856
Fakhfakh-Maala et al,[21] 2022	X	-	X	X	X	X	-	-	X	-	0.89
Feretzakis et al,[22] 2024	X	-	X	-	X	X	-	-	-	TS, age	0.715
Glicksberg et al,[23] 2024	X	-	-	X	X	X	GPT-4	-	-	-	0.88
Graham et al,[24] 2018	X	X	X	-	X	-	-	X	-	Age, arrival mode, TS	0.859
Handley et al,[25] 2015	X	X	-	X	X	-	-	-	-	-	0.86
Hong et al,[26] 2018	X	X	X	X	X	X	-	-	-	TS, age, retired status	0.874
Huo et al,[27] 2019	X	X	X	X	X	X	-	X	X	Arrival mode, HR, age	0.922
Kim et al,[28] 2022	X	X	X	X	X	X	-	-	-	TS, BP, GCS	0.933
King et al,[29] 2022	X	X	X	-	X	X	-	X	-	Age, arrival mode, PP	0.81
Kishore et al,[30] 2023	X	X	X	X	X	X	Extracted words	-	-	Age, arrival mode, PP	0.93
Mehra et al,[31] 2024	X	X	X	X	X	-	NLP	-	-	Age, arrival mode, "sepsis"	0.84
Monahan et al,[32] 2022	X	-	-	-	X	X	-	-	-	BP, RR and temperature	0.841
Park et al,[33] 2023	X	-	-	-	-	X	-	-	X	Age, RR, and temperature	0.75
Raita et al,[34] 2019	X	-	X	X	X	X	-	-	X	Age, arrival mode, BP	0.82
Rendell et al,[35] 2019	X	X	X	X	X	-	-	X	-	Arrival mode, age, nursing home	0.827
Ryu et al,[36] 2022	X	X	X	X	X	X	-	-	-	TS, weight, had ECG by triage	0.88
Sax et al,[37] 2023	X	-	-	-	X	X	MLP	X	X	-	0.87
Somanchi et al,[38] 2022	X	X	X	X	X	-	-	X	-	-	0.868
Sterling et al,[39] 2019	-	-	-	-	-	-	NLP	-	-	-	0.785
Tschoellitsch et al,[40] 2023	X	X	X	-	X	X	-	X	-	Age, mobility, suitcase brought	0.80

Continued on next page

TABLE 3. Continued

Reference, year	Demographics	Arrival time	Arrival mode	Presenting complaint	Triage score	Vital signs	Triage free-text	Past presentations or admissions	Medical history	Most important features	AUROC
Uhl et al,[41] 2022	X	X	X	X	X	-	-	-	-	Age, shock or trauma pathway	0.87
Yao et al,[42] 2020	X	X	X	-	X	X	-	-	-	-	0.86
Zhang et al,[43] 2017	X	X	X	-	X	-	NLP	X	X	Chest pain, shortness of breath	0.846
Zlotnik et al,[44] 2016	X	-	X	X	X	-	-	X	-	Shortness of breath, diabetes, TS	0.863

BP, blood pressure; ECG, echocardiogram; GCS, Glasgow Coma Scale; HR, heart rate; MH, medical history; PP, past presentation; RR, respiratory rate; TS, triage score.
[a]AUROC not reported in manuscript so accuracy performance reported instead.

range narrowed from 0.81 to 0.93. Of these models in set A who reported accuracy, values of 74%,[15] 79%,[41] 82.9%,[23] and 87.1% were reported.[30] This evidenced that models produced with good rigor and transparency have the potential to perform above the level of an experienced triage nurse. Regarding set A, the highest performing models were an ensemble including XGB and GPT-4 free text processing (AUROC, 0.88),[23] multilayer perceptrons (MLPs; AUROC, 0.858-0.87),[41,44] and stand-alone XGB models (AUROC, 0.81-0.93).[20,29,30] This finding of increased performance and robustness in articles with NLP, MLP, and XGB models is in slight contrast to previous reviews where logistic regression models are more commonly represented.[10,48] This is partially explained by their increasing prevalence and our review analyzing only the highest performing model from each articles, rather than all models developed.

Although 8 NNs in this review are simple MLPs, 1 study included an attention-based NN[17] and 1 included a convolutional NN,[42] with reported AUROC of 0.94 and 0.86, respectively. Although these types of NNs represent more advanced NN techniques, these studies failed to demonstrate temporal validity and had concerns regarding handling of missing data[17] and lack of basic demographic patient data.[42] These models have the potential to provide significant performance improvements and warrant further investigation.

To assess the performance of an admission prediction model in an ED, it is imperative to demonstrate prospective predictive ability. We define this as the ability to predict admission to hospital from triage data in patient data derived from a period outside the training and internal validation period. This is essential because health care is a dynamic system with constant fluctuations on hospital, population, and global levels, which can potentially alter model performance. For example, Duckworth et al[20] and King et al[29] compared development and testing data from separate periods and showed decreases in performance of 1% to 4%.[20,29] This occurred over an approximately 1.5-month and 5.5-month temporal validation set period.[20,29] This degree of drift over a short period implies that although long-term performance drift is unknown, it may become significant. This phenomenon, referred to as concept drift, is a commonly recognized problem during ML and AI model implementation.[49,50] In other domains, many models can identify concept drift and, in some cases, adapt to drift with ongoing tuning of hyperparameters.[49,50] This could be important for both rapidly changing health crises such as COVID-19 and major influenza seasons, or as hospital admission patterns change owing to changing demographics, hospital services, and out-of-hospital services such as hospital in the home. Unfortunately, no articles explored long-term concept drift or concept drift adaptation strategies. Admission prediction modeling would likely benefit from implementing innovations from other domains to quantify concept drift and concept drift adaptation.

Feature Selection and Importance

Feature selection is an important factor to consider when designing a model and implementation plan. Although some studies used specific feature selection methods, most models used all features in the database available. The importance of features and their contribution to model performance between studies was difficult to analyze owing to significant confounders such as population factors, hospital data and workflow, study methodology, and variable rigor. Furthermore, some included features demonstrated complex relationships with admission rates. For example, the analysis by Monahan et al[32] showed a parabolic relationship between many vital signs and admission frequency. They found both high and low values of heart rate, blood pressure, respiratory rate, and temperature were associated with increased rate of admission.[32]

Despite the features with the highest importance being used commonly, model performance remains variable between studies, even with similar features and model types. Therefore, we must consider other factors contributing to performance. For example, the number of features included in a model appears to correlate with improved performance. Models using a broad number of features, or a set of features with NLP, often outperformed models with more limited ranges of features. However, there are many confounders between articles limiting the ability to quantify the impact of increasing features. For example, presenting complaint can comprise a single categorical feature or in some articles comprise over 200 individual features for each possible presenting complaint. Therefore, analysis via simply the number of features can be misleading. However, within the 2 studies that analysed this relationship directly, AUROC increased by 6% and 11% when including increasing numbers of features or adding free-text analysis.[26,37] Therefore, there is emerging evidence that the inclusion of a broader set of features, possibly including free text, can lead to improved performance. Validation across a homogenous population or set of studies is required to confirm this relationship.

Missing Data and Exclusion

Handling of missing data and deriving exclusion criteria is a critical decision in model development, with significant implications for real-world implementation. Although some EDs have mandatory fields in the triage proforma, many do not, leading to patients with missing data. This is especially prevalent in free text boxes and early in a patient's hospital stay where vital signs may not be recorded. If a model is trained and tested on a data set excluding any case with missing data, the results may not reflect real-world conditions. Furthermore, inclusion or exclusion of patients with complex dispositions can limit external validity. These dispositions include patients who pass away in ED, did not wait, or were transferred to other hospital settings. Although many studies excluded these cases, this can reduce applicability. However, if too many of these cases are included in training, it may also skew the model and reduce accuracy overall. Considerations of workflow, system structure, and impact of including or excluding these cases are important to consider when trying to analyze the potential applicability of a model to clinical practice. Further studies should explore the impact of including and excluding these case types on model performance both in silico and during implementation.

Usability and Implementation

Of the studies meeting inclusion criteria, 22 were considered to have low useability in the clinical context. The primary reason for this is lack of prospective predictive ability and high concern for bias and applicability as per the PROBAST tool ratings. Of the remaining 9 articles, 1 had low accuracy,[15] 3 had less than 9 months of external validation data,[20,29,44] 2 lacked outcome definitions,[25,38] and 1 lacked description of handling of missing data.[25] This resulted in a moderate degree of usability for these 6 articles. The remaining 3 articles were considered to have good usability due to rigorous methodology and good accuracy.[23,30,41] However, they all lacked concept drift analysis or adaptation. Therefore, although XGB and NN models have shown some promise, further analysis of these models' performance over time would help determine whether these models are appropriate for implementation or require updating over time.

Beyond assessing concept drift analysis, testing in a clinical environment, or simulation in a near-clinical environment, would elucidate the utility of this modeling. Accuracy in the 80% to 90% range would generally not be considered adequate by physicians to dictate an individual patient's need for admission or discharge. It is possible that models may be able to provide a clinical aid in the case of patients being discharged despite the model strongly predicting admission. This may provide a clinical decision safety net, although the effect of such a system on clinical decision making, patient safety, and patient flow is debatable. However, with the degree of accuracy demonstrated, it is possible these models could positively contribute to patient flow planning. The study by King et al[29] was the only study in this review to report patient flow metrics directly from model development, plotting model-predicted beds against the observed required beds. Further assessment and reporting on degree of accuracy in bed prediction from triage data may help guide future implementation research and planning. If models can demonstrate adequate performance bed prediction, they could be implemented into the ED workflow as a live tool to provide predicted required beds based from the patients in the ED without a current disposition. These data could be made live, or via frequent reports, to patient flow managers and patient flow systems, which currently manage hospital patient flow planning. To analyze the impact of these models, commonly reported patient flow outcomes need to be measured including: ED length of stay, change in patient wait times, change in bed-block patterns, patient experience, user experience, and cost-benefit analysis.[4] Further analysis of performance over time, bed prediction accuracy, and ED patient flow measures are key areas in progressing from theoretical modeling to real-world implementation.

ML Developments

Advancements in ML and AI are rapidly changing the landscape of many industries, and health care is seeing a rapid expansion of research in this space. Within this review, although XGB and simple MLP NNs were the most common models, a more advanced attention-based NN,[17] a transformer network (GPT-4),[23] and a variety of other NLP techniques were used.[15,31,39,43] In the wider literature, transformer networks are becoming increasingly explored in the health care field after its original proposal in 2017,[51] particularly in analyzing large quantities of data and notes from health care databases, interpreting medical imaging investigations, DNA analysis, and molecular drug-drug interactions.[52] These networks with additional attention features may provide improved accuracy over simple ML techniques that have predominantly been used in admission prediction studies so far, although they have larger computational and energy demands. Attention-based NN and the transformer networks may have more promise in models analyzing increasingly large data sets, such as a model which continues to update its admission prediction decision over time as medical notes, further vital signs, laboratory tests, and medical imaging become available. Their ability to process these data is also dependant on the availability of these data over the course of a patient's ED stay. This can be due to limitations in the currently licenced software, delay in progress note creation or finalization, and even data that are still documented as a paper record. Typed medical notes may often be delayed from the time of patient review, depending on the workflow and clinical demands of the ED, delaying the availability of information for these systems. With emerging trials attempting to convert recorded ED medical reviews into real-time medical progress notes, this could significantly improve the possibilities for timely predictions of admissions.[53] Therefore, with emerging AI advancements, there are additional techniques and possibilities, which can improve admission prediction from both triage and throughout the ED stay. It may be appropriate for more computationally simple XGB models to be used for triage-based admission prediction owing to the narrower range of features and computational simplicity these models offer. However, the potential role for more complex models may become more apparent as more data become available and additional capabilities are desired from these models.

Limitations and Future Research

There are multiple limitations to address in the assessment and development of ML models predicting admission to hospital. First, use of the PROBAST and TRIPOD+AI quality appraisal tool, although completed according to protocol, they do not provide a quantitative assessment of performance. This approach was chosen to assess and address the heterogenicity of model development, analysis, and validation, given a meta-analysis would be unreliable. This subjectivity is compounded by the lack of clear gold standards in ML model development in this field, such as handling of missing data.

Heterogeneity is also created by variable populations between studies, even between different hospitals within a region, can create highly variable model performance. For example, Ryu et al[36] showed significant difference between multiple hospitals within the same state with the same model, with a change in AUROC from 0.84 to 0.94. Furthermore, when considering models trained at 1 hospital and tested at the same hospital compared with another hospital within the same region, AUROC ranged from 0.71 to 0.93.[36] This suggests population and hospital factors such as admission criteria and workflow can significantly impact model performance and its reporting. For example, for hospitals that use emergency short stay units, if patients who are admitted to this unit are included in admission, it can lead to significant improvement in performance from an AUROC of 0.86 to 0.93.[30] Although it is common in Australia to include admission to an emergency short stay units as inpatient admission, this may also increase the accuracy of a model.

It is also important to consider the outcome of inpatient admission as a complex and often subjective measure. It is often at the discretion of the admitting physician in consideration of medical risk, social factors, community supports, hospital policy, and hospital capacity. These physician, population, and hospital factors are significant confounders when analyzing model performance and make identifying the source of model accuracy more difficult. Prediction of model performance and clinical impact remains theoretical given no articles presented data derived from real-world implementation. Furthermore, given the lack of concept drift analysis, there are significant limitations in the confidence of a model to remain accurate over time. It remains unclear whether performance would be maintained in real-world application or whether concept drift adaptation is required as it is in other areas of ML implementation. When building and assessing ML models for admission prediction, consideration and reporting of feature availability, patient characteristics, hospital characteristics, and model development are crucial.

CONCLUSION

Overall, the models appraised in this review show in silico evidence that, when developed appropriately, ML can effectively predict admission from patient data commonly available at triage. XGB and MLP models are the most used ML models with an acceptable degree of accuracy, although accuracy can vary greatly between different populations and hospital settings. Commonly collected features such as age, triage category, arrival mode, and vital signs are among the most significant contributors to performance. A wider range of features also appears to improve performance, although this association requires further evidence. However, improved model performance may also be heavily dependent on local population and hospital characteristics. This challenges the notion of model performance being a product of mostly feature selection and model design. Rather, if ML predictive models become more prevalent in health care, local variations in workflow and system structures may become significant predictors of model performance. This may require development focusing on the nuances of local hospitals and health care systems.

Future research must strive to report model development, analysis, and findings in accordance with guidelines such as the PROBAST or the TRIPOD+AI criteria to ensure transparency and improve confidence in reported findings. Further research will hopefully help determine the optimal methodology for model development, explore the effectiveness of emerging AI techniques, analyze concept drift, and analyze the impact of real-world implementation on patient flow outcome measures.

POTENTIAL COMPETING INTERESTS

Dr Williams is supported by research grants from Australian Government Research Training Program (RTP) Stipend Scholarship and Noton Dale Medical Research Scholarship, via the University Of Notre Dame Fremantle, Australia. Dr Kanagasingam reports research grants from St John of God Midland Public Private Hospital. The other authors report no competing interests.

GRANT SUPPORT

This review was registered with PROSPERO (CRD42024580183) before analysis. The lead author (E.W.) received funding through an Australian Government Research Training Program Scholarship and the Noton-Dale Medical Research Scholarship at the University of Notre Dame, Australia. The funders had no role in study design, data collection, analysis, interpretation, or manuscript preparation.

SUPPLEMENTAL ONLINE MATERIAL

Supplemental material can be found online at https://www.mcpdigitalhealth.org/. Supplemental material attached to journal articles has not been edited, and the authors take responsibility for the accuracy of all data.

Abbreviations and Acronyms: **AI,** artificial intelligence; **AUROC,** area under the receiver operator characteristic curve; **ED,** emergency department; **ML,** machine learning; **MLP,** multilayer perceptron; **NLP,** natural language processing; **NN,** neural network; **XGB,** extreme gradient boosting

Correspondence: Address to Ethan L. Williams, MD, School of Medicine, The University of Notre Dame, 21 Henry Street, Fremantle, Western Australia 6160, Australia (ethan.williams1@my.nd.edu.au).

ORCID

Ethan L. Williams: https://orcid.org/0009-0003-7278-0739

REFERENCES

1. Mai Q, Aboagye-Sarfo P, Sanfilippo FM, Preen DB, Fatovich DM. Predicting the number of emergency department presentations in Western Australia: a population-based time series analysis. *Emerg Med Australas*. 2015;27(1):16-21. https://doi.org/10.1111/1742-6723.12344.
2. He J, Hou XY, Toloo S, Patrick JR, Fitz Gerald G. Demand for hospital emergency departments: a conceptual understanding. *World J Emerg Med*. 2011;2(4):253-261. https://doi.org/10.5847/wjem.j.1920-8642.2011.04.002.
3. Savioli G, Ceresa IF, Gri N, et al. Emergency department overcrowding: understanding the factors to find corresponding solutions. *J Pers Med*. 2022;12(2):279. https://doi.org/10.3390/jpm12020279.
4. Samadbeik M, Staib A, Boyle J, et al. Patient flow in emergency departments: a comprehensive umbrella review of solutions and challenges across the health system. *BMC Health Serv Res*. 2024;24(1):274. https://doi.org/10.1186/s12913-024-10725-6.
5. Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med*. 2019;170(1):W1-W33. https://doi.org/10.7326/M18-1377.
6. Brasen CL, Andersen ES, Madsen JB, et al. Machine learning in diagnostic support in medical emergency departments. *Sci Rep*. 2024;14(1):17889. https://doi.org/10.1038/s41598-024-66837-w.
7. Mueller B, Kinoshita T, Peebles A, Graber MA, Lee S. Artificial intelligence and machine learning in emergency medicine: a narrative review. *Acute Med Surg*. 2022;9(1):e740. https://doi.org/10.1002/ams2.740.
8. Petrella RJ. The AI future of emergency medicine. *Ann Emerg Med*. 2024;84(2):139-153. https://doi.org/10.1016/j.annemergmed.2024.01.031.
9. El-Bouri R, Taylor T, Youssef A, Zhu T, Clifton DA. Machine learning in patient flow: a review. *Prog Biomed Eng (Bristol)*. 2021;3(2):022002. https://doi.org/10.1088/2516-1091/abddc5.
10. Fernandes M, Vieira SM, Leite F, Palos C, Finkelstein S, Sousa JMC. Clinical decision support systems for triage in the emergency department using intelligent systems: a review. *Artif Intell Med*. 2020;102(1):101762. https://doi.org/10.1016/j.artmed.2019.101762.
11. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385(1):e078378. https://doi.org/10.1136/bmj-2023-078378.
12. Wolff RF, Moons KGM, Riley RD, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58. https://doi.org/10.7326/M18-1376.
13. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic

reviews. *Rev Esp Cardiol (Engl Ed)*. 2021;74(9):790-799. https://doi.org/10.1016/j.rec.2021.07.010.

14. Ahmed A, Ashour O, Ali H, Firouz M. An integrated optimization and machine learning approach to predict the admission status of emergency patients. *Expert Syst Appl*. 2022;202(1):117314. https://doi.org/10.1016/j.eswa.2022.117314.
15. Akhlaghi H, Freeman S, Vari C, et al. Machine learning in clinical practice: evaluation of an artificial intelligence tool after implementation. *Emerg Med Australas*. 2024;36(1):118-124. https://doi.org/10.1111/1742-6723.14325.
16. Araz OM, Olson D, Ramirez-Nafarrate A. Predictive analytics for hospital admissions from the emergency department using triage information. *Int J Prod Econ*. 2019;208(1):199-207. https://doi.org/10.1016/j.ijpe.2018.11.024.
17. Chen TL, Chen JC, Chang WH, Tsai W, Shih MC, Nabila AW. Imbalanced prediction of emergency department admission using natural language processing and deep neural network. *J Biomed Inform*. 2022;133(1):104171. https://doi.org/10.1016/j.jbi.2022.104171.
18. Cusidó J, Comalrena J, Alavi H, Llunas L. Predicting hospital admissions to reduce crowding in the emergency departments. *Appl Sci (Basel)*. 2022;12(21):10764. https://doi.org/10.3390/app122110764.
19. De Hond A, Raven W, Schinkelshoek L, et al. Machine learning for developing a prediction model of hospital admission of emergency department patients: hype or hope? *Int J Med Inform*. 2021;152(1):104496. https://doi.org/10.1016/j.ijmedinf.2021.104496.
20. Duckworth C, Chmiel FP, Burns DK, et al. Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19. *Sci Rep*. 2021;11(1):23017. https://doi.org/10.1038/s41598-021-02481-y.
21. Fakhfakh-Maala K, Ben-Othman S, Zgaya-Biau PH, Jourdan PL, Renard JM, Hammadi PS. Dynamic Dempster Multi-Layer Perceptron for the Prediction of Admission Patient in Emergency Department. *IEEE*. 2022;(1):190-195. https://doi.org/10.1109/SMC53654.2022.9945580.
22. Feretzakis G, Sakagianni A, Anastasiou A, et al. Integrating Shapley values into machine learning techniques for enhanced predictions of hospital admissions. *Appl Sci (Basel)*. 2024;14(13):5925. https://doi.org/10.3390/app14135925.
23. Glicksberg BS, Timsina P, Patel D, et al. Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room. *J Am Med Inform Assoc*. 2024;31(9):192-1928. https://doi.org/10.1093/jamia/ocae103.
24. Graham B, Bond R, Quinn M, Mulvenna M. Using data mining to predict hospital admissions from the emergency department. *IEEE Access*. 2018;6(1):10458-10469. https://doi.org/10.1109/access.2018.2808843.
25. Handly N, Thompson DA, Li JX, Chuirazzi DM, Venkat A. Evaluation of a hospital admission prediction model adding coded chief complaint data using neural network methodology. *Eur J Emerg Med*. 2015;22(2):87-91. https://doi.org/10.1097/mej.0000000000000126.
26. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. *PLoS One*. 2018;13(7):e0201016. https://doi.org/10.1371/journal.pone.0201016.
27. Huo Z, Sundararajhan H, Hurley NC, Haimovich A, Taylor RA, Mortazavi BJ. Sparse Embedding for Interpretable Hospital Admission Prediction. *Annu Int Conf IEEE Eng Med Biol Soc*. 2019;2019(1):3438-3441.
28. Kim E, Han KS, Cheong T, Lee SW, Eun J, Kim SJ. Analysis on benefits and costs of machine learning-based early hospitalization prediction. *IEEE Access*. 2022;10(1):32479-32493. https://doi.org/10.1109/access.2022.3160742.
29. King Z, Farrington J, Utley M, et al. Machine learning for real-time aggregated prediction of hospital admission for emergency patients. *NPJ Digit Med*. 2022;5(1):104. https://doi.org/10.1038/s41746-022-00649-y.
30. Kishore K, Braitberg G, Holmes NE, Bellomo R. Early prediction of hospital admission of emergency department patients. *Emerg Med Australas*. Aug 2023;35(4):572-588. https://doi.org/10.1111/1742-6723.14169.
31. Mehra T, Wekhof T, Keller DI. Additional value from free-text diagnoses in electronic health records: hybrid dictionary and machine learning classification study. *JMIR Med Inform*. 2024;12(1):e49007. https://doi.org/10.2196/49007.
32. Monahan AC, Feldman SS, Fitzgerald TP. Reducing crowding in emergency departments with early prediction of hospital admission of adult patients using biomarkers collected at triage: retrospective cohort study. *JMIR Bioinform Biotechnol*. 2022;3(1):e38845. https://doi.org/10.2196/38845.
33. Park SJ, Choi SH, Song DJ, et al. Improving triage accuracy of hospitalization and discharge decisions in the emergency department. *Signa Vitae*. 2023;19(5):75-90. https://doi.org/10.22514/sv.2023.015.
34. Raita Y, Goto T, Faridi MK, Brown DFM, Camargo CA Jr, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. *Crit Care*. 2019;23(1):64. https://doi.org/10.1186/s13054-019-2351-7.
35. Rendell K, Koprinska I, Kyme A, Ebker-White AA, Dinh MM. The Sydney Triage to Admission Risk Tool (START2) using machine learning techniques to support disposition decision-making. *Emerg Med Australas*. 2019;31(3):429-435. https://doi.org/10.1111/1742-6723.13199.
36. Ryu AJ, Romero-Brufau S, Qian R, et al. Assessing the generalizability of a clinical machine learning model across multiple emergency departments. *Mayo Clin Proc Innov Qual Outcomes*. 2022;6(3):193-199. https://doi.org/10.1016/j.mayocpiqo.2022.03.003.
37. Sax DR, Warton EM, Sofrygin O, et al. Automated analysis of unstructured clinical assessments improves emergency department triage performance: a retrospective deep learning analysis. *J Am Coll Emerg Physicians Open*. 2023;4(4):e13003. https://doi.org/10.1002/emp2.13003.
38. Somanchi S, Adjerid I, Gross R. To predict or not to predict: the case of the emergency department. *Prod Operat Manage*. 2022;31(2):799-818. https://doi.org/10.1111/poms.13580.
39. Sterling NW, Patzer RE, Di M, Schrager JD. Prediction of emergency department patient disposition based on natural language processing of triage notes. *Int J Med Inform*. 2019;129(1):184-188. https://doi.org/10.1016/j.ijmedinf.2019.06.008.
40. Tschoellitsch T, Krummenacker S, Dünser MW, Stöger R, Meier J. The value of the first clinical impression as assessed by 18 observations in patients presenting to the emergency department. *J Clin Med*. 2023;12(2):724. https://doi.org/10.3390/jcm12020724.
41. Uhl L, Augusto V, Lemaire V, et al. Progressive Prediction of Hospitalisation and Patient Disposition in the Emergency Department. *IEEE*. 2022:1719-1728. https://doi.org/10.1109/BigData55660.2022.10020777.
42. Yao LH, Leung KC, Hong JH, Tsai CL, Fu LC. A System for Predicting Hospital Admission at Emergency Department Based on Electronic Health Record Using Convolution Neural Network. *IEEE*. 2020:546-551. https://doi.org/10.1109/SMC42975.2020.9282952.
43. Zhang XY, Kim J, Patzer RE, Pitts SR, Patzer A, Schrager JD. Prediction of emergency department hospital admission based on natural language processing and neural networks. *Methods Inform Med*. 2017;56(5):377-389. https://doi.org/10.3414/me17-01-0024.
44. Zlotnik A, Alfaro MC, Pérez MC, Gallardo-Antolín A, Martínez JM. Building a decision support system for inpatient admission prediction with the manchester triage system and administrative check-in variables. *Comput Inform Nurs*. 2016;34(5):224-230. https://doi.org/10.1097/CIN.0000000000000230.
45. Trotzky D, Shopen N, Mosery J, et al. Real-time prediction of patient disposition and the impact of reporter confidence on mid-level triage accuracies: an observational study in Israel. *BMJ Open*. 2021;11(12):e050026. https://doi.org/10.1136/bmjopen-2021-050026.

46. Cameron A, Ireland AJ, McKay GA, Stark A, Lowe DJ. Predicting admission at triage: are nurses better than a simple objective score? *Emerg Med J*. 2017;34(1):2-7. https://doi.org/10.1136/emermed-2014-204455.
47. Holdgate A, Morris J, Fry M, Zecevic M. Accuracy of triage nurses in predicting patient disposition. *Emerg Med Australas*. 2007;19(4):341-345. https://doi.org/10.1111/j.1742-6723.2007.00996.x.
48. Larburu N, Azkue L, Kerexeta J. Predicting hospital ward admission from the emergency department: a systematic review. *J Pers Med*. 2023;13(5):849. https://doi.org/10.3390/jpm13050849.
49. Bayram F, Ahmed BS, Kassler A. From concept drift to model degradation: an overview on performance-aware drift detectors. *Knowledge Based Syst*. 2022;245(1):108632. https://doi.org/10.1016/j.knosys.2022.108632.
50. Hinder F, Vaquet V, Brinkrolf J, Hammer B. Model-based explanations of concept drift. *Neurocomputing*. 2023;555(1):126640. https://doi.org/10.1016/j.neucom.2023.126640.
51. Vaswani A. Attention is all you need. In: von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, eds. *Advances in Neural Information Processing Systems*. Curran Associates; 2017:6000-6010.
52. Nerella S, Bandyopadhyay S, Zhang J, et al. Transformers and large language models in healthcare: a review. *Artif Intell Med*. 2024;154(1):102900. https://doi.org/10.1016/j.artmed.2024.102900.
53. Tierney AA, Gayre G, Hoberman B, et al. Ambient artificial intelligence scribes to alleviate the burden of clinical documentation. *NEJM Catal Innov Care Deliv*. 2024;5(3):CAT.23.0404. https://doi.org/10.1056/CAT.23.0404.