Machine Learning Versus Usual Care for Diagnostic and Prognostic Prediction in the Emergency Department: A Systematic Review

Hashim Kareemi, MD[1], Christian Vaillancourt, MD, MSc[1,2], Hans Rosenberg, MD[1], Karine Fournier, MSI[3], and Krishan Yadav, MD, MSc[1,2]

A related article appears on page 263.

ABSTRACT

Objective: Having shown promise in other medical fields, we sought to determine whether machine learning (ML) models perform better than usual care in diagnostic and prognostic prediction for emergency department (ED) patients.

Methods: In this systematic review, we searched MEDLINE, Embase, Central, and CINAHL from inception to October 17, 2019. We included studies comparing diagnostic and prognostic prediction of ED patients by ML models to usual care methods (triage-based scores, clinical prediction tools, clinician judgment) using predictor variables readily available to ED clinicians. We extracted commonly reported performance metrics of model discrimination and classification. We used the PROBAST tool for risk of bias assessment (PROSPERO registration: CRD42020158129).

Results: The search yielded 1,656 unique records, of which 23 studies involving 16,274,647 patients were included. In all seven diagnostic studies, ML models outperformed usual care in all performance metrics. In six studies assessing in-hospital mortality, the best-performing ML models had better discrimination (area under the receiver operating characteristic curve [AUROC] =0.74–0.94) than any clinical decision tool (AUROC =0.68–0.81). In four studies assessing hospitalization, ML models had better discrimination (AUROC =0.80–0.83) than triage-based scores (AUROC =0.68–0.82). Clinical heterogeneity precluded meta-analysis. Most studies had high risk of bias due to lack of external validation, low event rates, and insufficient reporting of calibration.

Conclusions: Our review suggests that ML may have better prediction performance than usual care for ED patients with a variety of clinical presentations and outcomes. However, prediction model reporting guidelines should be followed to provide clinically applicable data. Interventional trials are needed to assess the impact of ML models on patient-centered outcomes.

[1]From the, Department of Emergency Medicine, University of Ottawa, Ottawa, Ontario; [2]the, Ottawa Hospital Research Institute, Ottawa, Ontario, Canada; [3]and the, Health Sciences Library, University of Ottawa, Ottawa, Ontario, Canada.
Received August 13, 2020; revision received October 6, 2020; accepted October 9, 2020.
The authors have no relevant financial information or potential conflicts to disclose.
Author contributions: HK helped formulate the research question with KY and CV, helped design the search strategy with KF, performed screening and full-text review with HR, performed supplemental and gray literature search, contacted authors for full texts, performed quality assessment, performed data extraction, created tables and figures, and wrote the manuscript with input from KY, CV, HR, and KF. CV helped formulate the research question; performed quality assessment; edited all drafts of manuscript, tables, and figures; and advised regarding methodologic considerations of review and included studies. HR performed screening and full-text review with HK, helped critically appraised systematic review results, and edited the manuscript. KF designed the search strategy, advised regarding technical considerations of review, and edited the manuscript. KY helped formulate the research question; adjudicated disagreements between HK and HR in screening and full-text review; performed quality assessment; edited all drafts of manuscript, tables, and figures; and advised regarding methodologic considerations of review and included studies. All authors reviewed and approved this final version of the manuscript.
Supervising Editor: Alice M. Mitchell, MD, MS.
Address for correspondence and reprints: Dr. Hashim Kareemi; e-mail: hkareemi@toh.ca.
ACADEMIC EMERGENCY MEDICINE 2021;28:184–196.

ISSN 1553-2712

doi: 10.1111/acem.14190

Optimal emergency department (ED) patient care depends on quick and accurate clinical decisions based on limited information and is becoming increasingly challenging. Over the past 20 years, ED visits in the United States have increased by approximately 30%,[1,2] contributing to increased crowding, costs, and delays in care.[3,4] These delays result in increased morbidity and mortality for ED patients.[5,6] Clinical prediction tools, such as the Canadian CT Head Rule and Quick Sequential Organ Function Assessment score, have been developed to support decision making under these demanding circumstances.[7,8] However, these tools remain limited to specific patient cohorts and clinical scenarios, and their development requires substantial time and resource investment.[9] Furthermore, there is growing evidence these tools do not always outperform unaided clinician judgment.[10,11]

Machine learning (ML) provides a novel approach to clinical prediction modeling. Where conventional statistical models rely on nonadaptive linear and logistic regression algorithms that use preprogrammed rules derived from highly specific clinical predictors, newer ML models utilize adaptable, nonparametric algorithms that can incorporate a greater breadth of complex predictors while maintaining predictive performance.[12] A major advantage of this approach lies in the ML model's ability to identify highly complex patterns across existing large data sets involving billions of data points (so-called "big data"). In supervised ML, models are trained on data that are "labeled" with the outcome to inform which combinations of predictors results in the most accurate prediction, thereby improving performance when tested with "unlabeled" input data. Supervised ML models have demonstrated diagnostic capability equal or superior to clinicians in several medical fields including medical imaging and dermatology.[13,14] ML models can also outperform conventional models in predicting clinical deterioration and need for intensive care unit (ICU) admission based on acute monitoring parameters of hospitalized patients.[15,16] Despite these promising studies, ML's potential for diagnostic and prognostic prediction remains largely unknown in emergency medicine.

We sought to compare the diagnostic and prognostic prediction performance of ML models using clinical variables to "usual care," including clinician judgment, conventional clinical prediction tools and triage-based scores, for patients presenting to the ED. If ML models can perform similarly to or better than usual care methods in retrospective analyses, this may provide justification for future interventional studies to assess the impact of ML models on clinical care and further incorporation of ML models into ED workflow.

METHODS

Study Design

We completed a systematic review of the literature with consideration of possible meta-analysis. Internal review board approval is not required for this type of study at our institution. The protocol was registered on PROSPERO (registration number CRD42020158129) and can be accessed at https://www.crd.york.ac.uk/prospero/display_record.php?ID = CRD42020158129. We reported our findings according to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines.[17]

Search Strategy

A health sciences librarian (KF) developed the search strategy for MEDLINE (OVID; see Data Supplement S1, Figure S1, available as supporting information in the online version of this paper, which is available at http://onlinelibrary.wiley.com/doi/10.1111/acem.14190/full) and then adapted it to the additional electronic databases (platform): Embase (OVID), Central (OVID), and CINAHL (EBSCOHost). A second librarian peer reviewed our search strategy according to the Peer Review of Electronic Search Strategies (PRESS) guidelines.[18] We searched all databases from date of inception to October 17, 2019, without date or language restrictions. We also hand searched references of included studies and relevant review articles. We conducted a gray literature search on December 1, 2019, by reviewing emergency medicine conference proceedings retrieved by our search, as well as by searching clinical trial registries (ClinicalTrials.gov, World Health Organization's International Clinical Trials Registry Platform) and Google Scholar. We contacted study authors regarding unpublished data for any potentially relevant abstracts.

Study Selection

We uploaded search strategy results into the Covidence systematic review software (Veritas Health Information, Melbourne, Australia) and removed duplicates electronically. Reviewers were trained in piloted screening procedures and extraction sheets to ensure

functionality. Two reviewers (HK and HR) independently performed screening and full-text review, with disagreement resolved by a third reviewer (KY). We included studies that compared one or more ML models to at least one method of usual care (e.g., clinician judgment, clinical decision rule, triage-based score) for diagnostic or prognostic prediction (e.g., mortality, ICU admission, or hospital length of stay) related to the index ED visit. A composite definition of "usual care" was used to capture the variation in diagnostic and prognostic approaches used for different clinical presentations and disease states, thereby including comparators with relevance to clinical practice and guidelines. Patient assessments must have occurred in the ED using input predictors that are readily available to an ED physician (e.g., triage vital signs, physical examination findings, and basic laboratory investigations). We excluded studies of ML models limited to test characteristics of specific diagnostic modalities (e.g., accuracy of medical imaging or electrocardiogram interpretation). We excluded studies involving only pediatric or prehospital patients as these utilize different predictors and usual care models than those used by emergency physicians working in adult academic EDs. While advanced logistic regression models could utilize functions found in newer ML models, most clinical prediction tools use conventional logistic regression models. To focus this review on newer ML models and distinguish them from other clinical prediction tools, we pragmatically excluded logistic regression from our definition of "usual care" unless they were part of a validated prediction model. Finally, we excluded nonprimary studies (i.e., letters, editorials, and review articles), case reports and series, studies for which no full-text article could be retrieved, and conference abstracts without a published full-text study or for which the authors did not respond to requests for a full-text reference.

Outcomes and Analysis

We extracted data according to previously published recommendations for systematic reviews of prediction models (PROBAST tool, CHARMS checklist, Debray guide).[19-21] The main outcome of interest was any prediction performance metric of model discrimination (e.g., area under the receiver operating characteristic curve [AUROC]), calibration (e.g., calibration plot slope), or classification (e.g., sensitivity, specificity). We present a narrative synthesis of our findings.

Risk of Bias Assessment

We used the Prediction Model Risk of Bias Assessment Tool (PROBAST) to assess risk of bias and applicability of the included ML models.[20] This tool uses 20 signaling questions to assess four realms (participants, predictors, outcome, and analysis) plus an overall judgment of risk of bias and applicability. Assessment of each study was performed independently by two of three reviewers (HK and either CV or KY). Any disagreements between reviewers in any of the individual realms prompted discussion and resolution by consensus.

RESULTS

We identified 1,656 unique records through our electronic search strategy, of which 96 were identified for full-text review. There was disagreement about five of 96 (5%) full-text articles for inclusion, all of which were ultimately excluded through consensus between reviewers, including one that required adjudication with the third reviewer. One author provided numerical data that had not been included in the original publication.[22] One additional study was retrieved in the gray literature search.[23] Twenty-three studies met full inclusion criteria (Figure 1).

Characteristics of the Included Studies

Study characteristics are described in Table 1. A total of 16,274,647 patients were assessed across 36 validation cohorts in which a unique outcome was tested. These cohorts involved internal (n = 30) or external (n = 6) validation. All included studies were observational cohort designs (17 retrospective; six prospective). Presenting complaints assessed in the studies were any chief complaint (n = 7), chest pain (n = 5), sepsis (n = 3), gastrointestinal bleed (n = 2), and a variety of others. A total of 72 ML models were tested, ranging from one to 17 models per study. The most commonly tested ML model types were (where n refers to number of studies that included one or more of that model type) neural network (n = 13), random forest (n = 11), gradient boosting (n = 8), and support vector machine (n = 5). Usual care methods included 20 clinical prediction rules and three triage-based scores. Six studies used unaided clinician judgment as the comparator, three of which were chest pain studies. The most common clinical outcomes assessed were clinical diagnosis of specific conditions (n = 8) such as sepsis and myocardial infarction, in-hospital mortality

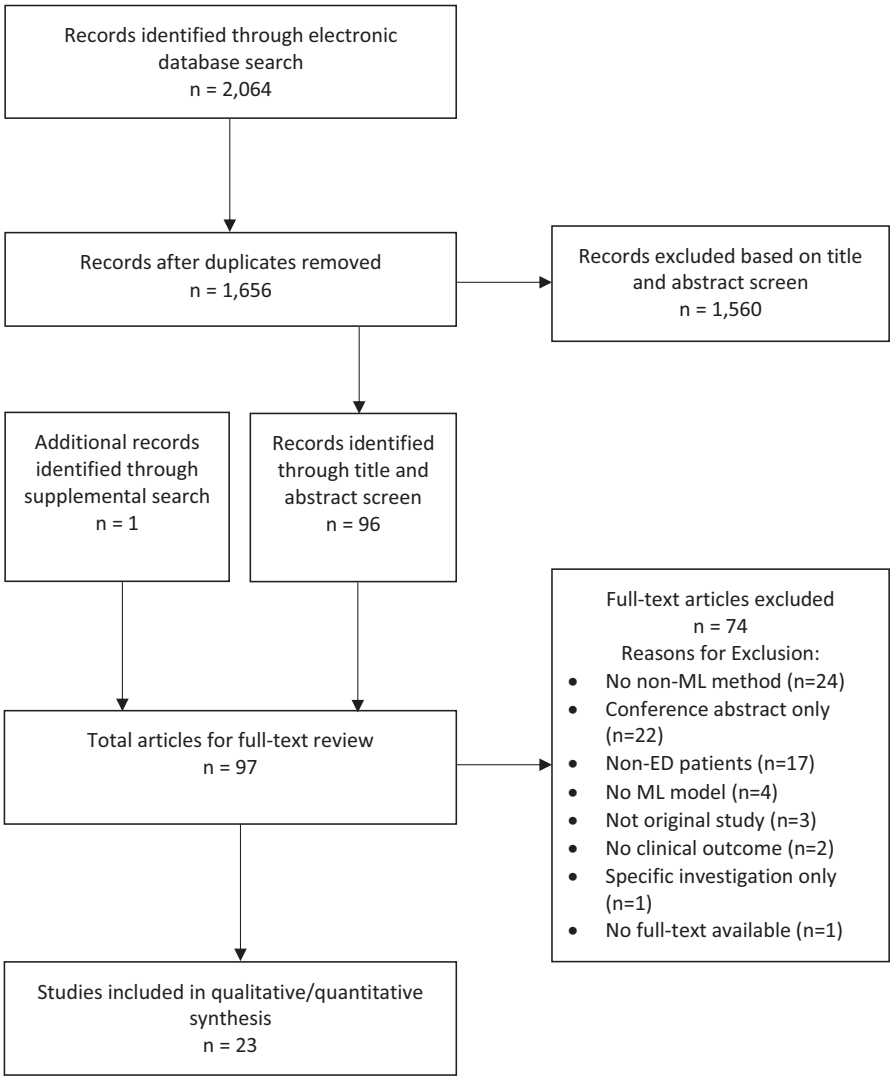

Figure 1. Study identification and selection. ML = machine learning.

(*n* = 7), other critical care outcomes (*n* = 6), and cardiac arrest or major adverse cardiac event (*n* = 3). "Critical care outcomes" was a variably defined composite outcome of ICU admission, need for therapeutic intervention, and/or in-hospital mortality. The number of participants with the outcome provides a measure of prevalence in the validation cohort. The events per variable refers to the ratio of the event rate in the derivation cohort to the number of candidate predictors used to develop the model.

Prediction Performance of ML and Usual Care

Machine learning models outperformed usual care methods for most diagnostic and prognostic predictions. For brevity, the performance metrics of only the best performing ML models and usual care methods are described in Table 2. Detailed performance metrics for all models are described in Data Supplement S1, Table S1. The best-performing ML models had better discrimination than any usual care method for all clinical outcomes, and better performance metrics in all external validation cohorts. In diagnostic studies, the ML models (AUROC =0.93, sensitivity =0.52–0.97, specificity =0.76–0.96, accuracy =74–96) outperformed usual care (AUROC =0.78, sensitivity =0.29–0.73, specificity =0.66–0.93, accuracy =65–84) in all performance metrics.[24-30] In three chest pain studies, ML models more accurately predicted a diagnosis of acute myocardial infarction than an emergency physician's judgment using the same variables.[24,26,27] A variety of prognostic outcomes were assessed. In six studies assessing in-hospital mortality, the best-performing ML models (AUROC =0.74–0.94) outperformed the best-performing clinical decision rules (AUROC =0.69–0.81) in discrimination

Table 1
Characteristics of the 23 Included Studies

First Author, Year	Country	Participant Data Collection	Validation Cohort Compared	Total Patients	% Male	Age (years)*	Patients With Outcome†	Events per Variable‡	ML Algorithms§	Usual Care Method§	Outcomes
All Patients Presenting to ED											
Delahanty, 2019	United States	Retrospective	Internal	2,759,529	42	48	54,661	976	GB	SOFA, SIRS, NEWS, MEWS, qSOFA	Diagnosis of sepsis, in-hospital mortality
Jang, 2019	South Korea	Retrospective	Internal	374,605	46	52	1,097	122	MLP, LSTM, Hybrid ANN, RF	MEWS	Cardiac arrest
Kwon, 2018	South Korea	Retrospective	Internal and external	10,967,518	50	50	150	11,384	MLP	MEWS, Korean Triage and Acuity Score	In-hospital mortality, critical care outcome, hospitalization
Levin, 2018	United States	Retrospective	Internal	172,726	44	47	46,525	4,652	RF	ESI	Critical care outcome, therapeutic intervention, hospitalization
Ong, 2012	Singapore	Prospective	Internal	925	62	61	43	2	SVM	MEWS	Cardiac arrest, in-hospital mortality
Raita, 2019	United States	Retrospective	Internal	135,470	57	46	2,782	259	LaR, RF, GB, ANN	ESI	Critical care outcome, hospitalization
Rendell, 2019	Australia	Retrospective	Internal	1,721,294	49	N/A	701,178	63,743	BN, DT, MLP, NB, kNN	Sydney Triage to Assessment Risk Tool	Hospitalization
Chest Pain											
Baxt, 1991	United States	Prospective	External	331	58	52	36	2	ANN	Clinician judgment	Diagnosis of myocardial infarction
Baxt, 1996	United States	Prospective	External	1,070	70	53	75	3	ANN	Clinician judgment	Diagnosis of myocardial infarction
Baxt, 2002	United States	Retrospective	Internal	2,204	40	53	361	9	ANN	ACI-TIPI, Goldmann algorithm	Diagnosis of acute ischemic heart disease
Kennedy, 1997	UK	Prospective	External	375	58	65	22	1	ANN	Clinician judgment	Diagnosis of myocardial infarction
Liu, 2014	Singapore	Prospective	External	702	66	60	29	1	RF	TIMI, MEWS	Major adverse cardiac event
Sepsis											
Chiew, 2019	Singapore	Retrospective	Internal	214	51	68	40	1	kNN, RF, AB, GB, SVM	qSOFA, NEWS, MEWS, Singapore ED Sepsis	Mortality
Perng, 2019	Taiwan	Retrospective	Internal	42,440	54	62	5,939	112	AE, CNN, PCA, ANN, RF, kNN, SVM, SoftMax	SIRS, qSOFA	Mortality

(Continued)

Table 1 (continued)

First Author, Year	Country	Participant Data Collection	Validation Cohort Compared	Total Patients	% Male	Age (years)*	Patients With Outcome†	Events per Variable‡	ML Algorithms§	Usual Care Method§	Outcomes
Taylor, 2016	United States, Scotland, England, Denmark, Singapore, New Zealand	Retrospective	Internal	5,278	46	66	50	13	RF	Mortality in ED Sepsis, Modified Rapid EM Score, CURB−65	Mortality
Gastrointestinal Bleeding											
Ayaru, 2015	UK	Retrospective	Internal	170	53	70	120	3	GB	BLEED, Strate score	Therapeutic intervention, severe bleeding, recurrent bleeding
Shung, 2019	United States	Retrospective	Internal and external	2,357	60	63	234	36	GB	Rockall, Glasgow-Blatchford score, AIMS65	Composite: therapeutic intervention or mortality
Syncope											
Falavigna, 2019	Italy	Retrospective	Internal	1,825	46	N/A	1,844	184	MLP	San Franscico syncope rule, OESIL	Composite: mortality, therapeutic intervention, intensive care unit admission, early readmission
Asthma/Chronic Obstructive Pulmonary Disease											
Goto, 2018	United States	Retrospective	Internal	3,206	40	52	128	75	LaR, RF, GB, ANN	ESI	Critical care outcome, hospitalization
Nonspecific Complaints											
Jenny, 2015	Switzerland	Retrospective	Internal	1,278	39	81	1,186	14	17 different models	Clinician judgment	Mortality, morbidity, acute infection
Laceration											
Lammers, 2003	United States	Prospective	Internal	1,142	N/A	29	82	3	ANN	Clinician judgment	Diagnosis of infection
Head Injury											
Molaei, 2016	United States	Retrospective	Internal	N/A	N/A	N/A	N/A	N/A	RF	Canadian CT head rule	Diagnosis of traumatic brain injury
Urinary Tract Infection											
Taylor, 2018	United States	Retrospective	Internal	80,387	32	53	18,284	1,828	RF, GB, AB, SVM, EN, ANN	Clinician judgment	Diagnosis of urinary tract infection

ML = machine learning.
*Presented age is the reported value from each study, whether median or mean. If the age was reported for separate cohorts, a weighted mean was calculated and presented.
†Refers to event rate in the validation cohort, whether internal or external. If multiple outcomes were reported, we present the first reported outcome.
‡Refers to event rate in the derivation cohort, where "variables" is the number of candidate predictors evaluated for potential inclusion in ML model.
§ML algorithms: AB = adaptive boosting; AE = autoencoder; ANN = artificial neural network; BN = Bayesian network; CNN = convolutional neural network; DT = decision tree; EN = elastic net; GB = gradient boosting; kNN = k-nearest neighbor; LaR = lasso regression; LSTM = long-short-term memory; MLP = multilayer perceptron; NB = naïve Bayes; PCA = principal component analysis; RF = random forest; SVM = support vector machine.
¶Usual care: AIMS65 = (albumin, INR, altered mental status, systolic blood pressure, age ≥65 years). BLEED = (persistent bleeding, low systolic blood pressure, elevated prothrombin time, erratic mental status, unstable comorbid disease); CURB-65 = (confusion, urea nitrogen, respiratory rate, blood pressure, ≥65 years); ESI = emergency severity index; MEWS = modified early warning score; NEWS = national early warning score; OESIL = Osservatorio Epidemiologico sulla Sincope nel Lazio; qSOFA = quick sequential organ failure assessment; SIRS = systemic inflammatory response syndrome; SOFA = sequential organ failure assessment; TIMI = thrombolysis in myocardial infarction.

Table 2
Prediction Performance of Best Performing ML Models and Usual Care from the 23 Included Studies

First Author, Year	Patient Population	Development Cohort	Clinical Outcome Category	ML Type*	AUROC (95% CI)	Acc (95% CI)	Sens (95% CI)	Spec (95% CI)	Usual Care Type*	AUROC (95% CI)	Acc (95% CI)	Sens (95% CI)	Spec (95% CI)
Diagnosis													
Baxt, 1991	Chest pain	External validation	Diagnosis—MI	ANN	—	96	0.97 (0.97–0.98)	0.96 (0.96–0.96)	Judgment	—	84	0.78 (0.77–0.83)	0.85 (0.84–0.86)
Baxt, 1996	Chest pain	External validation	Diagnosis—MI	ANN	—	96	0.96 (0.91–1.00)	0.96 (0.95–0.97)	Judgment	—	81	0.73 (0.63–0.83)	0.81 (0.79–0.84)
Baxt, 2002	Chest pain	Internal validation	Diagnosis—ischemic heart disease	ANN—Jackknife	—	87	0.88 (0.85–0.91)	0.86 (0.85–0.88)	Goldman	—	71	0.67 (0.63–0.72)	0.71 (0.69–0.73)
Kennedy, 1997	Chest pain	External validation	Diagnosis—MI	ANN	—	74 (74–74)	0.52 (0.52–0.53)	0.80 (0.80–0.80)	Judgment	—	65 (65–65)	0.29 (0.28–0.29)	0.77 (0.77–0.77)
Delahanty, 2019	Sepsis	Internal validation	Diagnosis—sepsis	GB	0.93	—	0.68	0.96	SOFA	0.78	—	0.49	0.93
Molaei, 2016	TBI	Internal validation	Diagnosis—TBI	RF	—	79	0.82	0.76	CCHR	—	65	0.64	0.66
Taylor, 2018	UTI	Internal validation	Diagnosis—UTI	GB (all variables)	—	84 (83–84)	0.80 (0.79–0.81)	0.85 (0.84–0.85)	Judgment	—	75 (74–76)	0.41 (0.40–0.43)	0.85 (0.84–0.85)
Prognosis —In-hospital Mortality													
Chiew, 2019	Sepsis	Internal validation	In-hospital mortality	SVM	—	—	0.63	—	NEWS	—	—	0.69	—
Delahanty, 2019	Sepsis	Internal validation	In-hospital mortality	GB	0.89	—	0.56	0.96	SOFA	0.78	—	0.53	0.92
Perng, 2019	Sepsis	Internal validation	In-hospital mortality	CNN +SoftMax	0.94 (0.94–0.94)	87			qSOFA	0.74 (0.73–0.74)	67		
Perng, 2019	Sepsis	Internal validation	In-hospital mortality	CNN +SoftMax	0.92 (0.92–0.92)	87			qSOFA	0.68 (0.67–0.69)	67		
Taylor, 2016	Sepsis	Internal validation	In-hospital mortality	RF	0.86 (0.82–0.90)	—	—	—	CURB–65	0.73 (0.67–0.80)	—	—	—
Kwon, 2018	Any	Internal validation	In-hospital mortality	MLP	0.94 (0.94–0.94)	—	—	—	MEWS	0.81 (0.81–0.81)	—	—	—
Kwon, 2018	Any	External validation	In-hospital mortality	MLP	0.92	—	—	—	KTAS	0.80	—	—	—
Ong, 2012	Any (high acuity)	Internal validation	In-hospital mortality	SVM	0.74	—	0.70	0.74	MEWS	0.69	—	0.74	0.56
Prognosis—Critical Care Outcome													
Goto, 2018	Asthma, COPD	Internal validation	Critical care outcome	GB	0.80	—	0.79	0.68	ESI	0.68	—	0.53	0.81
Kwon, 2018	Any	Internal validation	Critical care outcome	MLP	0.89 (0.89–0.90)	—	—	—	KTAS	0.80 (0.80–080)	—	—	—
Levin, 2018	Any	Internal validation	Critical care outcome	RF	0.83	—	—	—	ESI	0.79	—	—	—
Raita, 2019	Any	Internal validation	Critical care outcome	ANN	0.86 (0.85–0.87)	—	0.80 (0.77–0.83)	0.76 (0.73–0.78)	ESI	0.74 (0.72–0.75)	—	0.50 (0.47–0.53)	0.86 (0.82–0.87)
Shung, 2019	Upper GIB	Internal validation	Critical care outcome	GB	0.91 (0.90–0.93)	—	—	—	GBS	0.88 (0.86–0.90)	—	—	—
Shung, 2019	Upper GIB	External validation	Critical care outcome	GB	0.90 (0.87–0.93)	—	—	—	GBS	0.87 (0.84–0.91)	—	—	—
Prognosis—Hospital Admission													
Goto, 2018	Asthma, COPD	Internal validation	Hospital admission	RF	0.83	—	0.75	0.76	ESI	0.64	—	0.33	0.84
Kwon, 2018	Any	Internal validation	Hospital admission	MLP	0.80 (0.80–0.80)	—	—	—	KTAS	0.68 (0.68–0.68)	—	—	—
Raita, 2019	Any	Internal validation	Hospital admission	ANN	0.82 (0.82–0.83)	—	0.79 (0.78–0.80)	0.71 (0.69–0.72)	ESI	0.69 (0.68–0.69)	—	0.87 (0.86–0.87)	0.42 (0.39–0.43)
Rendell, 2019	Any	Internal validation	Hospital admission	kNN	0.83 (0.83–0.83)	75 (75–75)	—	—	START	0.82 (0.82–0.82)	74 (74–74)	—	—
Prognosis—Other													
Ayaru, 2015	Lower GIB	Internal validation	Recurrent bleeding	GB	—	88	0.67	0.91	BLEED	—	64	0.24	0.75
Ayaru, 2015	Lower GIB	Internal validation	Therapeutic intervention	GB	—	88	0.80	0.89	BLEED	—	68	0.28	0.76
Ayaru, 2015	Lower GIB	Internal validation	Severe bleeding	GB	—	78	0.73	0.80	BLEED	—	63	0.33	0.79
Falavigna, 2019	Syncope	Internal validation	Severe short-term outcomes	ANN	0.90	—	1.00 (0.83–1.00)	0.79 (0.72–0.85)	SFSR	0.82	—	0.88 (0.62–0.98)	0.76 (0.67–0.83)

(Continued)

Table 2 (continued)

First Author, Year	Patient Population	Development Cohort	Clinical Outcome Category	ML Type*	AUROC (95% CI)	Acc (95% CI)	Sens (95% CI)	Spec (95% CI)	Usual Care Type*	AUROC (95% CI)	Acc (95% CI)	Sens (95% CI)	Spec (95% CI)
Jang, 2019	Any	Internal validation	Cardiac arrest	Hybrid ANN	0.94 (0.93–0.94)	—	0.94 (0.93–0.95)	0.75 (0.75–0.75)	MEWS	0.89 (0.88–0.89)	—	0.88 (0.87–0.89)	0.75 (0.75–0.75)
Jenny, 2015	Nonspecific complaints	Internal validation	All-cause mortality	Median of 17 models	0.82	—	—	—	Judgment	0.67	—	—	—
Jenny, 2015	Nonspecific complaints	Internal validation	Acute morbidity	Median of 17 models	0.80	—	—	—	Judgment	0.65	—	—	—
Jenny, 2015	Nonspecific complaints	Internal validation	Acute infection	Median of 17 models	0.71	—	—	—	Judgment	0.60	—	—	—
Lammers, 2003	Laceration	Internal validation	Wound infection	ANN	—	—	0.70	0.77	Judgment	—	—	0.54	0.78
Liu, 2014	Chest pain	External validation	MACE	RF (top 3 variables)	0.81 (0.72–0.91)	—	0.83 (0.69–0.97)	0.63 (0.60–0.67)	TIMI	0.64 (0.53–0.75)	—	—	—
Ong, 2012	Any (high acuity)	Internal validation	Cardiac arrest	SVM	0.78	—	0.81	0.72	MEWS	0.68	—	0.74	0.54

COPD = chronic obstructive pulmonary disease; GIB = gastrointestinal bleeding; MI = myocardial infarction; TBI = traumatic brain injury; UTI = urinary tract infection.
*ML algorithms: ANN = artificial neural network; BN = Bayesian network; CNN = convolutional neural network; GB = gradient boosting; kNN = k-nearest neighbor; MLP = multilayer perceptron; RF = random forest; SVM = support vector machine.
†Usual care: BLEED = (persistent bleeding, low systolic blood pressure, elevated prothrombin time, erratic mental status, unstable comorbid disease); CCHR = Canadian CT Head Rule; CURB-65 = (confusion, urea nitrogen, respiratory rate, blood pressure, ≥65 years); ESI = Emergency Severity Index; GBS = Glasgow-Blatchford score; KTAS = Korean triage and acuity score; MEWS = modified early warning score; NEWS = national early warning score; qSOFA = quick sequential organ failure assessment; SFSR = San Francisco syncope rule; SIRS = systemic inflammatory response syndrome; SOFA = sequential organ failure assessment; START = Sydney triage to assessment risk tool; TIMI = thrombolysis in myocardial infarction.

metrics.[23,28,31-34] One of these studies used area under precision-recall curve (AUPRC) instead of AUROC and showed that a gradient-boosting ML model (AUPRC =0.35) outperformed five clinical decision rules (AUPRC =0.21–0.29) for suspected sepsis patients.[31] ML models (AUROC =0.80–0.91) had better discrimination than usual care methods (AUROC =0.64–0.88) in the prediction of critical care outcomes and hospitalization, including four studies comparing performance to triage-based scores for undifferentiated ED patients presenting with any complaint.[22,33,35,36] There were three comparisons in which sensitivity[31,34,36] and three in which specificity[35-37] of the usual care model was higher than that of the best-performing ML model. However, ML models had better discrimination performance in all of these cases except for one study that did not measure discrimination.[37] In that study, ML had greater sensitivity whereas clinician judgment had greater specificity. In all cases, the overall accuracy of prediction at a given threshold was better for the ML models.

Clinical Heterogeneity

There was significant clinical heterogeneity between studies (see Data Supplement S1, Tables S2 and S3, for detailed study descriptions). ML models were trained differently in terms of the sampling procedure for the internal validation cohort and predictor selection. Patient selection and outcome criteria often differed across studies, and even when they aligned, the reported prediction measures often did not. Given these differences, we determined that a meta-analysis was not appropriate.

Quality Assessment and Risk of Bias

Quality assessment of the included studies using the PROBAST tool is described in Table 3, with details of individual criteria included in Data Supplement S1, Table S4. Overall, participants and predictors were well defined and selected appropriately, with 18 (78%) and 20 (87%) of 23 included studies judged to be at low risk of bias in these realms, respectively. In studies where there was high risk of bias, patients were inappropriately excluded or predictors were defined or assessed differently between participants. The majority of outcomes had inconsistent or unclear definitions, with only 15 (65%) studies having low risk of bias in this category. In several studies, the predictors were themselves components of the outcome definition, or the outcome was likely determined with knowledge of

Table 3
Risk of Bias and Applicability Assessment of Included Studies Using PROBAST Tool

	Risk of Bias				Applicability			Overall	
Author, Year	Participants	Predictors	Outcome	Analysis	Participants	Predictors	Outcome	Risk of Bias	Applicability
Ayaru, 2015	–	+	?	–	+	+	+	–	+
Baxt, 1991	–	+	–	–	+	+	?	–	?
Baxt, 1996	+	+	–	–	+	+	+	–	+
Baxt, 2002	+	+	–	–	+	+	+	–	+
Chiew, 2019	–	+	+	–	+	+	+	–	+
Delahanty, 2019	+	+	–	–	+	+	+	–	+
Falavigna, 2019	+	–	–	–	+	+	+	–	+
Goto, 2018	+	+	+	–	+	+	+	–	+
Jang, 2019	+	+	+	–	+	+	+	–	+
Jenny, 2015	+	–	+	–	+	+	+	–	+
Kennedy, 1997	+	+	–	–	+	+	+	–	+
Kwon, 2018	+	+	+	–	+	+	+	–	+
Lammers, 2003	+	–	+	–	+	+	+	–	+
Levin, 2018	+	+	+	–	+	+	+	–	+
Liu, 2014	–	+	+	–	+	+	+	–	+
Molaei, 2016	+	+	?	–	+	+	+	–	+
Ong, 2012	–	+	+	–	+	+	+	–	+
Perng, 2019	+	+	+	–	+	+	+	–	+
Raita, 2019	+	+	+	+	+	+	+	–*	+
Rendell, 2019	+	+	+	–	+	+	+	–	+
Shung, 2019	+	+	+	+	+	+	+	+	+
Taylor, 2018	+	+	+	+	+	+	+	–*	+
Taylor, 2016	+	+	+	–	+	+	+	–	+

+ = favorable judgment (low risk of bias or high applicability); – = unfavorable judgment (high risk of bias or low applicability); ? = uncertain judgment (not enough information provided)
*Downgraded due to lack of external validation.

the predictors (i.e., lack of blinding). Analysis was the most poorly reported section, with only three (13%) studies at low risk of bias. Two of these studies were subsequently downgraded due to lack of external validation. Most studies did not evaluate model calibration, instead reporting only classification measures such as sensitivity or specificity. Studies often did not report on how they handled continuous variables, missing data, or complexities in the data such as patient censoring and competing risks. Patients with missing data were often excluded (i.e., "case–control analysis"), and a description of which patients had missing data was infrequently provided.

Overall high risk of bias in the included studies precluded sensitivity analysis of low risk of bias studies. The use of the TRIPOD reporting guidelines appeared to mitigate many of these deficiencies.[38] Of the three studies that referenced TRIPOD, one had overall low risk of bias,[39] and two had issues with only one signaling question each.[30,36] Despite issues with risk of bias, applicability of the studies to our research question remained high. All but one study clearly used appropriate predictors for their ML models that would be available upon ED assessment and subsequently predicted relevant outcomes.

DISCUSSION

We systematically reviewed studies comparing ML models to usual care for diagnostic and prognostic prediction in the ED setting. Overall, ML models outperformed usual care in most diagnostic and prognostic predictions across all 23 included studies. In particular, the best-performing ML models had better discrimination abilities (AUROC) than any usual care method for all tested outcomes. However, methodologic and reporting issues in the included studies introduced substantial risk of bias. Significant clinical

heterogeneity between studies also precluded meta-analysis. Proceeding to interpret and apply these results to clinical practice must therefore be done cautiously.

Discrimination is defined as a model's ability to separate participants with a predicted outcome from those without it, and it is commonly measured using the concordance (C) statistic or AUROC curve.[40] A principal finding of our study is that ML models had better discrimination performance than usual care. Other systematic reviews have shown similar results, including one comparing mortality prediction for acute upper gastrointestinal bleeding, with all but one of 22 comparisons showing significantly higher discrimination scores for ML models than clinical prediction rules.[41] In a systematic review of neurosurgical studies, ML models performed better or similarly to clinician judgment in 47 of 50 (94%) studies assessing diagnosis, presurgical planning, and outcome prediction.[42] It is important to consider that AUROC itself has limited application in particular analyses. A recent study proposed using the AUPRC instead of AUROC to assess discrimination, because it incorporates prevalence and may thus be more informative in imbalanced data sets.[43] In the two studies that reported them in our review, ML models had higher AUPRC values than usual care.[31,33]

Calibration is defined as the agreement between a model's predicted outcomes and actual observed outcomes and can be quantified using a calibration plot slope, the ratio of observed to expected/predicted outcomes, or the Hosmer-Lemeshow test of fit.[21,40] Unlike discrimination, calibration was reported in only one included study; thus it could not be adequately assessed. Two systematic reviews have demonstrated that authors rarely perform or report calibration measures,[40,44] suggesting that this is a persistent issue in prediction modeling research.

Classification metrics such as sensitivity and specificity assess model performance at particular thresholds. Reclassification parameters, such as the net reclassification index, similarly provide a measure of "movement" between risk categories based on these thresholds. However, these thresholds often have unsubstantiated clinical relevance, thus introducing bias.[20] The so-called "threshold-free" measures of discrimination and calibration are therefore preferred for overall prediction model assessment.[43] When discrimination measures were not reported, we found one study where the sensitivity of the ML model was higher but specificity lower compared to usual care,[37] a similar finding to the review by Senders et al.[42] of neurosurgical studies. In our study, heterogeneity and poor reporting of thresholds made interpretation of classification measures difficult, which appears to be another persistent issue in prediction modeling research.[14]

In addition to performance metrics, the methodology of prediction model development and validation is important to evaluate for potential sources of bias. Models are "trained" using a derivation cohort and then "tested" on a validation cohort either internally or externally. External validation allows generalizability of the results and is considered the most rigorous form of prediction assessment,[20,40] yet it is infrequently used in prediction modeling studies.[14,44] Only six studies in our review used an external validation cohort when comparing ML models to usual care,[24,26,27,33,39,45] suggesting that the majority of included studies suffered from a significant risk of bias. Another commonly encountered source of bias is an inadequate number of participants with the outcome of interest.[40,44] The PROBAST guidelines recommend a minimum of 10 "events" or participants with the outcome per "variable" or candidate predictors for model derivation studies and 100 participants with the outcome for validation studies.[46] In our review, seven studies did not meet these criteria, risking overfitting or underfitting of the ML models.

Given the methodologic issues cited, bias may have influenced the results found among the included studies. Christodoulou et al.[44] demonstrated that ML and logistic regression models performed similarly in clinical prediction when only including studies with low risk of bias, but that ML outperformed logistic regression models in studies with high risk of bias. This issue also raises the question of publication bias, which is largely ignored in data analytics research and as such may unduly influence our perception of the benefits of ML and artificial intelligence in general.[47]

Our study has several strengths. Our electronic search strategy was developed by a health sciences librarian; was peer reviewed; and had no restriction by language, date, or particular ED presentation. This makes our results relevant not only to ED clinicians but also to consultants from other specialties who must assess and risk stratify these patients in the ED. Our search was also not limited to a particular date, and while this risked including older ML models that were built using smaller databases and were prone to overfitting, it allowed us to thoroughly document the

evolution of ML for clinical prediction, highlighting the improvements achieved and challenges that remain with more sophisticated modeling. We included clinician judgment as a comparator, an important addition as it may perform as well as conventional clinical prediction tools in many scenarios.[10,11] We used the PROBAST tool, which assesses specific and relevant criteria for prediction model development that are not addressed by other quality assessment tools. We also used previously published guidelines for prediction modeling research to direct our study development (Debray guide) and data extraction (CHARMS).[19,21]

Our study has several implications for clinical practice and future research. As numerous studies including our own have demonstrated, the quality of reporting for studies on ML prediction models requires improvement. Many of the reporting deficiencies identified by PROBAST and CHARMS in the constituent studies of our review were components of the TRIPOD guidelines, and we therefore recommend the use of TRIPOD to ensure adequate reporting of primary studies.[38] Development of an extension of these guidelines for ML studies (TRIPOD-AI) was announced in April 2019, but are not yet published. Reporting guidelines for systematic reviews of ML prediction models would also mitigate the limitations of the PROBAST guidelines. Furthermore, we advocate for future studies to investigate the effects of ML models on clinical outcomes using interventional trial designs. As noted in previous studies, there are several barriers to implementing ML models in real-time clinical practice as well as ensuring their uptake and proper use by clinicians.[48] Only by extending our work on predictive performance into clinical efficacy can we justify utilizing ML in clinical practice. Until then, it is important for practicing ED clinicians to understand that ML has the potential to provide more accurate diagnostic and prognostic prediction than currently available clinical tools.

LIMITATIONS

Our study has limitations that warrant mentioning. First, clinical heterogeneity of the included studies precluded meta-analysis. As discussed, this is likely a result of inconsistent reporting among prediction model research, and it has been similarly encountered in several other systematic reviews.[41,42,48] Second, our study was not designed to assess the effect of ML models on clinical outcomes themselves. Other reviews that included interventional studies found very few.[42,48] Accordingly, we chose to focus on prediction performance to allow for consistency in search criteria and interpretation of results. It is important to consider that many of the included clinical prediction tools themselves have limited applicability to clinical practice, and so comparatively superior performance by an ML model does not necessarily imply ideal or "best" practice. Third, the PROBAST tool limited our risk-of-bias assessment to the ML models being derived or validated and thus did not capture the execution and analysis of usual care methods. While other reviews have modified tools such as QUADAS-2 and QUIPS for diagnostic test accuracy and prognostic risk studies respectively,[41,48] these modifications are not validated and fail to capture important considerations of prediction model development. Several of the PROBAST criteria, in particular, the nine criteria for model analysis, were highly specific and sometimes difficult to extrapolate to ML model development, resulting in many studies being considered at high risk of bias.

CONCLUSIONS

In conclusion, our systematic review found that machine learning models appear to have better diagnostic and prognostic prediction performance compared to usual care for ED patients with a variety of presentations. Reporting guidelines must be developed and followed to ensure uniform reporting and comparison of studies assessing machine learning models clinical prediction performance. Future research should investigate the clinical impacts of machine learning models in interventional trials.

Ms. Angela Marcantonio assisted with formatting and submission of this manuscript for publication. Dr. Wei Chen assessed the results and helped conclude that a meta-analysis would not be feasible. Ms. Majela Guzman peer reviewed the electronic search strategy using the PRESS guidelines.

REFERENCES

1. Nourjah P. National Hospital Ambulatory Medical Care Survey: 1997 Emergency Department Summary. Advance

Data From Vital and Health Statistics. Hyattsville, MD: National Center for Health Statistics, 1999:304.
2. Rui P, Kang K. National Hospital Ambulatory Medical Care Survey: 2017 Emergency Department Summary Tables. Hyattsville, MD: National Center for Health Statistics, 2017.
3. Emergency Medicine Practice Committee. Emergency Department Crowding: High Impact Solutions. Irving, TX: American College of Emergency Physicians, 2016.
4. Gaieski DF, Agarwal AK, Mikkelsen ME, et al. The impact of ED crowding on early interventions and mortality in patients with severe sepsis. Am J Emerg Med 2017;35:953–60.
5. Chalfin DB, Trzeciak S, Likourezos A, Baumann BM, Dellinger RP; DELAY-ED study group. Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit. Crit Care Med 2007;35:1477-83.
6. Rathore SS, Curtis JP, Chen J, et al. Association of door-to-balloon time and mortality in patients admitted to hospital with ST elevation myocardial infarction: national cohort study. BMJ 2009;338:b1807.
7. Stiell IG, Wells GA, Vandemheen K, et al. The Canadian CT Head Rule for patients with minor head injury. Lancet 2001;357:1391-6.
8. Seymour CW, Liu VX, Iwashyna TJ, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016;315:762-74.
9. Stiell IG, Wells GA. Methodologic standards for the development of clinical decision rules in emergency medicine. Ann Emerg Med 1999;33:437-47.
10. Sanders S, Doust J, Glasziou P. A systematic review of studies comparing diagnostic clinical prediction rules with clinical judgment. PLoS One 2015;10:e0128233.
11. Schriger DL, Elder JW, Cooper RJ. Structured clinical decision aids are seldom compared with subjective physician judgment, and are seldom superior. Ann Emerg Med 2017;70:338-44.e3.
12. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med 2019;380:1347-58.
13. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115-8.
14. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digital Health 2019;1:e271-e297.
15. Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Crit Care Med 2016;44:368-74.
16. Wellner B, Grand J, Canzone E, et al. Predicting unplanned transfers to the intensive care unit: a machine learning approach leveraging diverse clinical elements. JMIR Med Inform 2017;5:e45.
17. Moher D. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Medicine 2009;6:e1000097.
18. McGowan J, Sampson M, Salzwegel DM, Cogo E, Foerster M, Lefebvre C. PRESS - Peer Review of Electronic Search Strategies: 2015 Guideline Explanation and Elaboration (Press E&E). Ottawa: CADTH, 2016.
19. Moons KG, de Groot JA, Bouwmeester W, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. PLoS Medicine 2014;11:e1001744.
20. 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;170:51-8.
21. Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460.
22. Levin S, Toerper M, Hamrock E, et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the Emergency Severity Index. Ann Emerg Med 2018;71:565-74.e2.
23. Perng JW, Kao IH, Kung CT, Hung SC, Lai YH, Su CM. Mortality prediction of septic patients in the emergency department based on machine learning. J Clin Med 2019;8:1906.
24. Baxt WG. Use of an artificial neural network for the diagnosis of myocardial infarction. Ann Intern Med 1991;115:843-8.
25. Baxt WG, Shofer FS, Sites FD, Hollander JE. A neural network aid for the early diagnosis of cardiac ischemia in patients presenting to the emergency department with chest pain. Ann Emerg Med 2002;40:575-83.
26. Baxt WG, Skora J. Prospective validation of artificial neural network trained to identify acute myocardial infarction. Lancet 1996;347:12-5.
27. Kennedy RL, Harrison RF, Burton AM, et al. An artificial neural network system for diagnosis of acute myocardial infarction (AMI) in the accident and emergency department: evaluation and comparison with serum myoglobin measurements. Comput Methods Prog Biomed 1997;52:93-103.
28. Delahanty RJ, Alvarez J, Flynn LM, Sherwin RL, Jones SS. Development and evaluation of a machine learning model for the early identification of patients at risk for sepsis. Ann Emerg Med 2019;73:334-44.

29. Molaei S, Korley FK, Reza Soroushmehr SM, et al. A machine learning based approach for identifying traumatic brain injury patients for whom a head CT scan can be avoided. Ann Int Conf IEEE Eng Med Biol Soc 2016;2016:2258-61.
30. Taylor RA, Moore CL, Cheung KH, Brandt C. Predicting urinary tract infections in the emergency department with machine learning. PLoS One 2018;13:e0194085.
31. Chiew CJ, Liu N, Tagami T, Wong TH, Koh ZX, Ong ME. Heart rate variability based machine learning models for risk prediction of suspected sepsis patients in the emergency department. Medicine 2019;98:e14197.
32. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach. Acad Emerg Med 2016;23:269-78.
33. Kwon JM, Lee Y, Lee Y, Lee S, Park H, Park J. Validation of deep-learning-based triage and acuity score using a large national dataset. PLoS One 2018;13:e0205836.
34. Ong ME, Lee Ng CH, Goh K, et al. Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. Crit Care 2012;16:R108.
35. Goto T, Camargo CA Jr, Faridi MK, Yun BJ, Hasegawa K. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. Am J Emerg Med 2018;36:1650-4.
36. Raita Y, Goto T, Faridi MK, Brown DF, Camargo CA Jr, Hasegawa K. Emergency department triage prediction of clinical outcomes using machine learning models. Crit Care 2019;23:64.
37. Lammers RL, Hudson DL, Seaman ME. Prediction of traumatic wound infection with a neural network-derived decision model. Am J Emerg Med 2003;21:1-7.
38. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med 2015;162:55-63.
39. Shung DL, Au B, Taylor RA, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. Gastroenterology 2020;158:160-7.
40. Bouwmeester W, Zuithoff NP, Mallett S, et al. Reporting and methods in clinical prediction research: a systematic review. PLoS Medicine 2012;9:1-12.
41. Shung D, Simonov M, Gentry M, Au B, Laine L. Machine learning to predict outcomes in patients with acute gastrointestinal bleeding: a systematic review. Dig Dis Sci 2019;64:2078-87.
42. Senders JT, Arnaout O, Karhade AV, et al. Natural and artificial intelligence in neurosurgery: a systematic review. Neurosurgery 2018;83:181-92.
43. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 2015;10:e0118432.
44. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22.
45. Liu N, Koh ZX, Goh J, et al. Prediction of adverse cardiac events in emergency department patients with chest pain using machine learning for variable selection. BMC Med Inform Decis Mak 2014;14:75.
46. Moons KG, 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:W1-33.
47. Boulesteix AL, Stierle V, Hapfelmeier A. Publication bias in methodological computational research. Cancer Inform 2015;14:11-9.
48. Fleuren LM, Klausch TL, Zwager CL, et al. Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy. Intensive Care Med 2020;46:383-400.

Supporting Information

The following supporting information is available in the online version of this paper available at http://onlinelibrary.wiley.com/doi/10.1111/acem.14190/full

Data Supplement S1. Supplemental material.