Infectious diseases now 55 (2025) 105090

Contents lists available at ScienceDirect

Infectious Diseases Now

journal homepage: www.sciencedirect.com/journal/infectious-diseases-now

Original article

Evaluating the impact of artificial intelligence in antimicrobial stewardship: a comparative meta-analysis with traditional risk scoring systems

Antonio Pinto [a], Flavia Pennisi [b,a,*], Giovanni Emanuele Ricciardi [a,b], Carlo Signorelli [a], Vincenza Gianfredi [c]

[a] Faculty of Medicine, University Vita-Salute San Raffaele, Milan, Italy
[b] National Program in One Health Approaches to Infectious Diseases and Life Science Research, Department of Public Health, Experimental and Forensic Medicine, University of Pavia, Pavia 27100, Italy
[c] Department of Biomedical Sciences for Health, University of Milan, Via Pascal 36, 20133 Milan, Italy

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Antimicrobial resistance
Antimicrobial stewardship
Meta-analysis

ABSTRACT

Objectives: The growing challenge of antimicrobial resistance (AMR) has underscored the urgent need for robust antimicrobial stewardship programs (AMS). Artificial intelligence (AI) and machine learning (ML) have emerged as promising tools to support enhanced decision-making in AMS. This systematic review and meta-analysis aims to evaluate the impact of AI in AMS and compare its effectiveness with traditional risk systems.

Methods: PubMed/MEDLINE, Scopus, EMBASE, and Web of Science were searched to identify studies published up to July 2024. Any studies that evaluated the use of AI/ML in AMS compared with conventional decision-making approaches were eligible. Outcomes of interested were predictive performance metrics and diagnostic accuracy. The meta-estimate was performed pooling standardized mean difference, and effect size (ES) measured as Cohen's d with a 95% confidence interval (CI). The risk of bias was assessed using the QUADAS-AI tool.

Results: Out of 3,458 studies, 27 were included, demonstrating that ML models outperform traditional methods in terms of sensitivity [1.93 (0.48–3.39) $p = 0.009$], and negative predictive value [1.66 (0.86–2.46), $p < 0.001$] but not in terms of area under the curve, accuracy, specificity, positive predictive value, when random effect models were applied.

Conclusions: Our results revealed that ML tools offer promising enhancements to traditional AMS strategies. However, high heterogeneity, inconsistent results between fixed and random effect models, and limited use of external validation in retrieved studies raise concerns about the generalizability of the findings. Furthermore, the lack of representation from outpatient and pediatric settings highlights a critical equity gap in the application of these technologies.

1. Introduction

Antimicrobial resistance (AMR) represents one of the most critical public health challenges globally, requiring innovative strategies to optimize antibiotic use and improve clinical outcomes. AMR undermines the efficacy of antimicrobial therapies, exacerbating the severity and incidence of infections while significantly increasing healthcare costs [1]. In 2021, AMR caused over 1.14 million deaths globally, with related infections raising this figure to 4.71 million. Projections indicate that, without effective interventions, AMR could be responsible for as many as 8.22 million deaths annually by 2050 [2].

Antimicrobial stewardship (AMS) is pivotal in curbing inappropriate antibiotic use and controlling resistant strains. First introduced in 1974 by McGowan and Finland [3], AMS aims to optimize the selection, dosage, and duration of antimicrobial therapies to improve clinical outcomes, reduce healthcare costs, and combat AMR [4]. AMS addresses antibiotic use across human, animal, and environmental health, yet 30–50 % of hospital use remains inappropriate, fueling resistance and highlighting the need for stronger interventions [5].

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) within AMS have drawn attention, promising to revolutionize clinical practice with data-driven insights.

* Corresponding author at: National Program in One Health Approaches to Infectious Diseases and Life Science Research, Department of Public Health, Experimental and Forensic Medicine, University of Pavia, Pavia 27100, Italy.
E-mail address: flavia.pennisi01@universitadipavia.it (F. Pennisi).

https://doi.org/10.1016/j.idnow.2025.105090
Received 10 February 2025; Received in revised form 14 March 2025; Accepted 12 May 2025
Available online 14 May 2025

Unlike traditional methods such as logistic regression – which relies on parametric and linearity assumptions, often limiting their capacity to handle complex, high-dimensional datasets – ML algorithms flexibly model nonlinear relationships and integrate diverse data sources. ML's ability to process large data volumes and identify complex patterns has sparked interest in its potential to improve predictive accuracy and address AMR in real time [6]. Regulatory agencies, including the Food and Drug Administration (FDA) and the European Medicines Agency (EMA), have begun exploring the integration of ML-based technologies into healthcare [7], spanning applications such as medical imaging [8], disease diagnosis, and patient management [9]. In the field of infectious diseases, ML has demonstrated utility in areas such as diagnostic accuracy, prognostic modeling, and treatment decision-making, paving the way for novel approaches to AMR mitigation [10]. This systematic review aims to evaluate and compare predictive performance of ML algorithms with logistic regression in the context of AMS. By systematically analyzing studies that utilize these approaches, the review seeks to identify differences in accuracy, sensitivity, specificity, and overall utility for predicting outcomes related to antibiotic use and stewardship interventions.

2. Methods

2.1. Protocol and registration

This systematic review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [11]. The protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO), registration number [PROSPERO ID CRD42024567640].

2.2. Literature search strategy

A comprehensive literature search was performed across PubMed/MEDLINE, Scopus, EMBASE, and Web of Science (WoS) in July 2024. This systematic review seeks to answer the following question: "Is AI more effective than traditional models in AMS for reducing AMR?". Traditional statistical methods were defined as linear regression, generalized linear models (*e.g.*, logistic regression), Cox proportional hazards regression, parametric survival models of all types, and simple non-model-based approaches (*e.g.*, calculating the observed risk within a subgroup). Other methodologies were classified as ML techniques, including but not limited to random forests, XGBoost, and naive Bayes. The search strategy incorporated two key elements: artificial intelligence (and synonyms) and antimicrobial stewardship or antimicrobial resistance (and synonyms). Selected keywords, including MeSH terms and Title/Abstract terms, were combined using the Boolean operators "AND" and "OR." The search strategy was initially developed for PubMed/MEDLINE and subsequently adapted for Scopus, EMBASE, and WoS. The search strategies for each database are detailed in **Table S1**. Additionally, forward and backward citation searches were performed. No temporal filter was applied.

2.3. Eligibility criteria

Eligibility criteria were established following the Population, Intervention, Comparison, Outcome, and Study design (PICOS) framework [12]. Studies were included if they involved patients, both inpatients and outpatients, requiring antimicrobial prescriptions in relevant settings and employed both traditional and ML models within AMS. Only studies that developed and, when applicable, validated ML models aimed at enhancing AMS practices were considered. A key inclusion criterion was that studies had to report performance metrics for both ML models and traditional models, allowing for a direct comparison.

We included original observational studies, either prospective or retrospective, published in English in peer-reviewed journals, which involved patients – both in inpatient and outpatient settings – requiring antimicrobial prescriptions and applied ML models within the context of AMS. To be eligible, studies had to compare ML models with traditional statistical approaches, such as logistic regression, and report at least one performance metric enabling direct comparison between the two methods (*e.g.*, area under the curve (AUC), accuracy, sensitivity, specificity, positive predictive value, and negative predictive value). Only studies that developed and, when applicable, validated ML models aimed at supporting AMS practices were considered. We excluded conference abstracts, editorials, reviews, case reports, studies not directly related to AMS, and those that did not provide comparative performance data between ML and traditional models.

A comprehensive description of the inclusion criteria, aligned with the PICOS framework, can be found in Table 1.

2.4. Study selection

Search results were imported into Mendeley for organization and removal of duplicates. Titles and abstracts were independently screened in a blinded manner by two reviewers utilizing Rayyan software [13]. Full-text articles were subsequently obtained and assessed blindly by the same two reviewers based on the established eligibility criteria. Disagreements at any stage of the selection process were independently resolved through consensus between the two reviewers. If consensus was not reached, a third reviewer was consulted to arbitrate the final decision. A PRISMA flowchart was created to document the study inclusion process, detailing the number of articles retained at each stage of screening and the reasons for exclusion during the full-text review.

2.5. Data extraction

Data extraction was conducted in a blind way by reviewers using a standardized template in Microsoft Excel (Microsoft Excel® for Microsoft 365 MSO, Redmond, WA, USA, 2019). The spreadsheet was initially piloted on three randomly selected articles to improve consistency and agreement among the authors, as previously performed [14]. Extracted data encompassed the following: study characteristics (author name, year of publication, country, and study design), population characteristics (demographic information, sample size, hospital setting, type of infection, and clinical environment), details of ML model (methods employed, training data sets, number of features, and data sources, including clinical and/or laboratory data), and outcomes of ML and traditional models (predictors, performance validation, and clinical results). Discrepancies in data extraction were addressed through discussion or by consulting a third reviewer.

Table 1
Inclusion criteria according to population, intervention, comparison, outcome, and study design (PICOS) guidelines.

	Inclusion criteria
Population (P)	Patients who require antimicrobial prescriptions in inpatient or outpatient settings
Intervention (I)	Artificial intelligence or machine learning tools used in antimicrobial stewardship
Comparison (C)	Traditional clinical decision-making processes without the aid of artificial intelligence
Outcome (O)	Main outcome: measure of the effectiveness of machine learning algorithms to support antimicrobial stewardship programs: AUC, c-Statistic, accuracy, sensitivity, specificity, PPV, NPV, and F1-score Secondary outcome: clinical outcomes and improvement in patient health outcomes
Study design (S)	Observational, including cohort studies, prospective or retrospective
Time filter	No temporal filter

AUC: Area under the curve; PPV: positive predictive value; NPV: negative predictive value.

2.6. Risk of bias assessment

The risk of bias in the studies included was independently and blindly assessed by two authors using the QUADAS-AI tool [15]. It is an extension of the QUADAS-2 and QUADAS-C5 tools, specifically designed to assess the risk of bias and applicability in AI-centered diagnostic test accuracy studies. It addresses AI-specific biases such as algorithmic bias and evaluates key domains: patient selection, focusing on the representativeness and quality of included data; index test, assessing the execution and interpretation of the diagnostic test; reference standard, ensuring the appropriateness of the comparison test; and validation type, emphasizing the need for both internal and external validation to avoid overfitting. Discrepancies were resolved through discussion or by consulting a third reviewer. This evaluation ensures the reliability and validity of findings.

2.7. Data synthesis and meta-analysis

Data synthesis was organized into tables to provide an overview of study characteristics and findings. In the meta-analysis, only performance measures with data from at least three distinct studies were included. Quantitative analyses were conducted on performance measurements related to validation, where available. A quantitative synthesis of the data was planned, if feasible, using fixed-effects and random-effects models to assess the AUC, accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of traditional and ML models. The effect size (ES) was calculated based on the mean and standard deviation (SD) or differences between means and SD, and the sample size provided for each study. In the current meta-analysis, the pooled ES was expressed as a standardized mean difference and measured as Cohen's d with a 95 % CI [16]. This metric is commonly defined as minor (d = 0.2), intermediate (d = 0.5), and substantial (d = 0.8) [17]. An I^2 test was conducted to assess the heterogeneity among the studies included: high if I^2 values exceeded 75 %, moderate for values ranging between 50 % and 75 %, low for values between 25 % and 50 %, and no heterogeneity if values were below 25 % [18]. Both graphical evaluation of the Funnel plot and Egger's regression asymmetry test were employed to assess potential publication bias, with statistical significance set at $p < 0.10$ [19]. For the sensitivity analysis, ML models were categorized based on functional similarities, including their data processing methods, underlying mechanisms, and common applications. The identified groups were as follows:

- Decision Trees, including Random Forest (RF), Decision Tree (DT), Classification and Regression Trees (CART), and J48 (C4.5). These types of models are a set of decision trees that make predictions by combining the results of multiple trained trees on random subsets of data and characteristics, improving accuracy, and reducing overfitting.
- Boosted Models, including Extreme Gradient Boosting (XGB), Gradient Boosted Decision Trees (GBDT), Categorical Boosting (CatBoost), Adaptive Boosting (AdaBoost), Boosted Decision Trees (Boosted DT), Boosted Logistic Regression (Boosted LR), and Gradient Boosting Machine (GBM). These models combine multiple weak decision trees in sequence, with each new tree correcting the errors of the previous one. This process optimizes predictions and reduces bias while minimizing the increase in variance.
- Neural Networks were divided into two categories: generic networks, such as Artificial Neural Networks (ANN), Neural Networks (NN), Backpropagation Neural Networks (Backpropagation NN), and Neural Networks with SHapley Additive exPlanations (NN with SHAP); and recurrent networks, including Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bidirectional LSTM), Recurrent Neural Networks (RNN), and Gated Recurrent Unit (GRU). These are models composed of layers of interconnected artificial neurons that transform data through adaptive weights and activation functions, learning complex patterns through iterative error optimization.
- Support Vector Machines (SVMs), including standard Support Vector Machines (SVM), SVM with Radial Basis Function kernel (SVM with RBF kernel), Sequential Minimal Optimization (SMO), Linear Support Vector Machines (Linear SVM), Polynomial Support Vector Machines (Polynomial SVM), and SVM C-Support Vector. These methods classify data by identifying the line or plane that best separates them into categories, maximizing the distance between these categories. For more complex data, mathematical techniques such as kernels are used to find an effective separation.

To address the heterogeneity among the studies included and better explore the different applications of ML, a secondary sensitivity analysis was conducted by stratifying studies into five groups (**Table S2**) based on their primary objectives:

- Diagnostic Microbiology and Resistance Detection (*e.g.*, using MALDI-TOF data to detect carbapenem-resistant pathogens) [39,43,44,46,58];
- Prediction of Culture Results and Bacteremia (*e.g.*, predicting blood culture outcomes from electronic health records) [34,40,41,45,48];
- Risk Stratification and Prediction of Multidrug Resistance (*e.g.*, predicting colonization or infection by multidrug-resistant organisms) [35,36,47,59,60];
- Optimization of Prescriptions and Clinical Decision Support (*e.g.*, identifying inappropriate antibiotic use and guiding stewardship interventions) [37,49–51,53–56];
- Epidemiological Surveillance, Outbreak Prediction and other specific application [38,42,52,57].

3. Results

3.1. Literature search

A total of 3,458 studies were identified through searches in PubMed/MEDLINE (n = 992), Scopus (n = 1,134), Embase (n = 705), and Web of Science (n = 627) databases. Following the initial removal of duplicates (n = 1,825), a total of 1,633 studies underwent screening based on title and abstract. Subsequently, 1,589 studies were eliminated due to non-original content and focus on different topics, resulting in 44 studies deemed eligible for inclusion. Three articles lacked full-text availability. Following full-text assessment, 14 studies were excluded, including 27 studies. The most common reason for exclusion was wrong outcome (n = 7) [20–26], followed by conference abstracts (n = 5) [27–31] and wrong study designs (n = 2) [32,33]. The study selection process is visually represented in Fig. 1.

3.2. Descriptive characteristics of included studies

The systematic review comprised 27 studies, encompassing a total of about 550 thousand patients. All studies were published after the year 2000. The geographic distribution is enough concentrated (Fig. 2), with the highest contribution from the United States (n = 7 studies) [34–40], Spain (n = 3) [41–43], Taiwan (n = 3) [44–46], and China (n = 2) [47,48]. Most studies (62.9 %, n = 17) [36,37,41–46,48–56] were monocentric (Table 2). Bloodstream infections were the predominant area of research focus, comprising 33.3 % (n = 9) of total studies [34,37,40,41,43,46,48,50,55]. Respiratory tract infections accounted for 25.9 % (n = 7) of the studies [37,46,47,53,56–58], followed by urinary tract infections (18.5 %, n = 5) [46,50,51,53,54]. Most studies (81.5 %, n = 22) [34–39,41–43,46–53,55,57–60] reported data on inpatient (I) populations, while no study only reported outpatient data (O). A further 11.1 % (n = 3) of studies [40,45,56] reported data on emergency patients (E), while only one included both inpatient and outpatient data (I/O) [54], and one did not specify this information (NA)

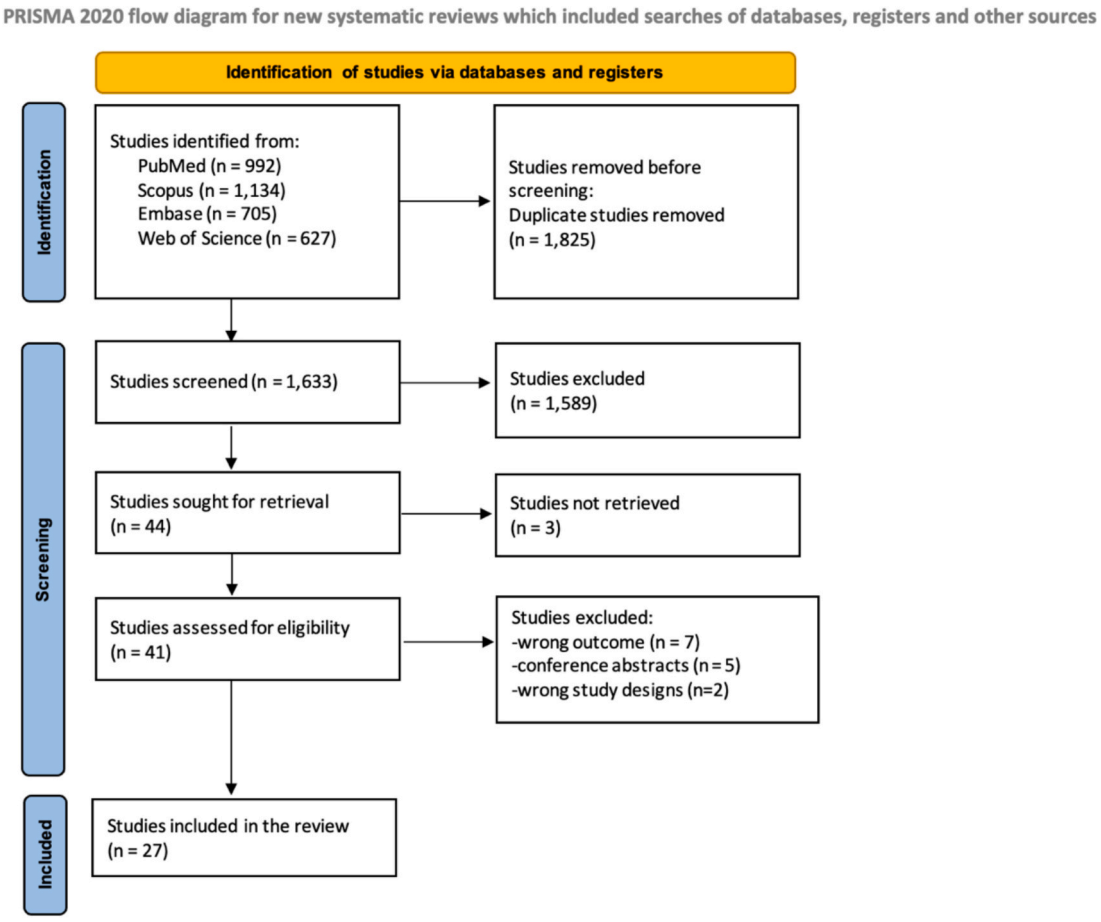


Fig. 1. Flow diagram depicting the selection process.

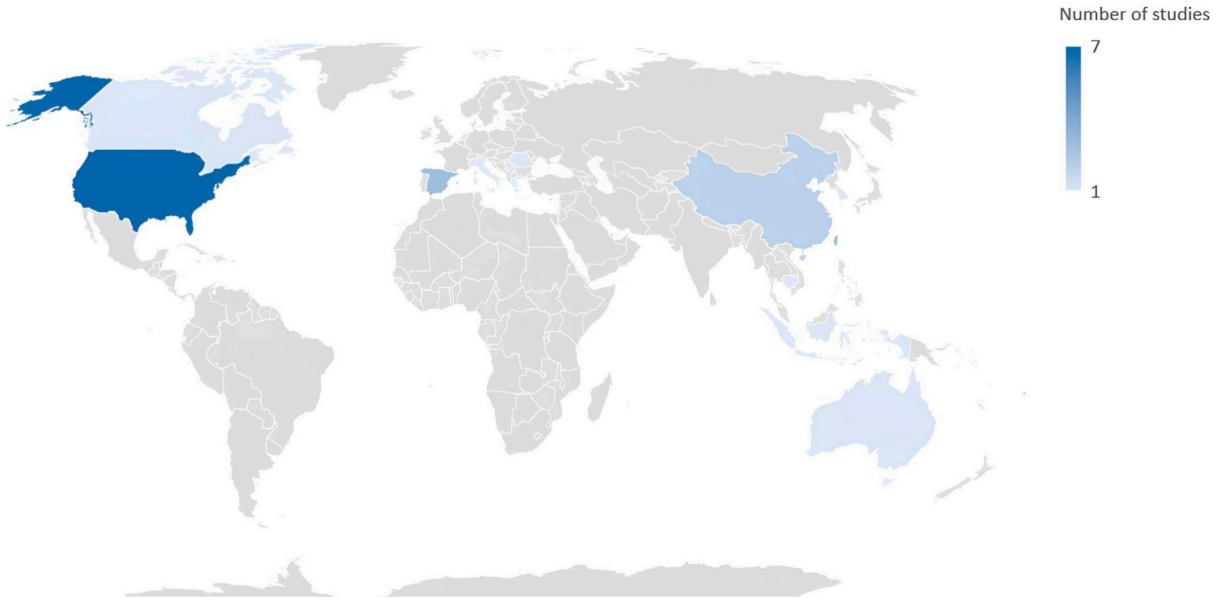


Fig. 2. Geographical distribution of included studies.

[44].

Of the 27 studies included, only three [38,47,51] used external validation and two did not specify this information (NA) [43,60]. The number of features used in the studies varied greatly from a minimum of 8 features [39] to a maximum of 23,067 [53], with a mean of 1,257. Calibration was performed in only four models [34,35,50,55].

3.3. Characteristics of features of included studies

Fig. 3 illustrates the distribution of key features analyzed in the

Table 2
Descriptive characteristics of included studies.

Author, year (Ref)	Study type	Country	No. of centers	Data time frame	Study objective	Target population (demographic data)	No. of patients	HS	Infection site	TDS / Validation	Cal.	Features
Ananda-Rajah M–R, 2017 [57]	RCS	Australia	3	NA	To develop an expert system for electronic surveillance of IMD that combines natural language processing of CT reports, microbiology, and antifungal drug data to improve prediction of IMD	Patients from hematology-oncology departments, with a focus on those at high risk of IMD, such as those undergoing hematopoietic stem cell transplantation	123; 64 with IMD and 59 controls	I	Respiratory	NA / Internal	No	NA
Beaudoin M, 2016 [49]	PCS	Canada	1	Feb-Nov 2012; Nov-Dec 2013	To evaluate the ability of the algorithm to discover rules for identifying inappropriate prescriptions of TZP	Patients monitored by APSS who received at least one prescription of TZP at the Centre Hospitalier Universitaire de Sherbrooke	421	I	NA	Yes / Internal	No	NA
Bhavani S, 2020 [34]	RCS	USA	2	2007–2018	To develop and validate ML models to predict blood culture results at the time of order using routine EHR data	Hospitalized adult patients who received a blood culture	Over 250,000 blood culture days across both centers	I	Bloodstream	Yes / Internal	Yes	NA
Brown D-G, 2023 [35]	PCS	USA	5	2017–2020	To derive and validate a clinical prediction rule for identifying USA international travelers at risk of acquiring ESBL-PE during travel	USA international travelers who were ESBL-negative before travel and provided stool samples before and after travel	528 travelers	I	Gastrointestinal	Yes / Internal	Yes	27
Bystritsky R-J, 2020 [50]	RCS	UAE	1	Dec 2015-Aug 2017	To predict whether antibiotic therapy required stewardship intervention on any given day compared to the criterion standard of note left by the antimicrobial stewardship team in the patient's chart	Patients hospitalized who received at least one antimicrobial from a list of those routinely tracked by the ASP at University of California, San Francisco Medical Centre	9,651; split randomly into derivation (80 %) and validation (20 %) data sets	I	Bloodstream, UTI	Yes / Internal	Yes	56
Çağlayan Ç, 2022 [36]	RCS	USA	1	2017–2018	To build a robust predictive analytics framework that produces reliable and evidence-based predictions with high sensitivity, ensuring timely detection of MDRO colonization, and high specificity, preventing inefficient use of limited resources	Patients admitted to a surgical or medical ICU	3,958 patients; 17.59 % MDRO, 13.03 % VRE, 1.45 % CRE, 7.47 % MRSA	I	NA	Yes / Internal	No	26
de Vries S, 2022 [51]	RCS	Netherlands	1	Jan 2017-Dec 2018	To report on the design and evaluation of a CDSS to predict UTI before the urine culture results are available	Inpatients of the UMC Utrecht	906 cultures from 810 patients	I	UTI	Yes / External	No	36
Eickelberg G, 2020 [52]	RCS	Israel	1	2011–2012	To identify ICU patients with low risk of bacterial infection as candidates for earlier EAT discontinuation	ICU adult patients are patients suspected of having a community-acquired bacterial infection	10,290 (12,232 ICU encounters); split into a training and test set following a 70/30 split	I	NA	Yes / Internal	No	NA
Feretzakis G, 2020 [53]	RCS	Greece	1	Jan 2017-Dec 2018	To compare the performance of eight ML algorithms to assess	ICU patients in a public tertiary hospital	345	I	Respiratory, UTI, mucocutaneous, wound	Yes / Internal	No	23,067

(*continued on next page*)

Table 2 (*continued*)

Author, year (Ref)	Study type	Country	No. of centers	Data time frame	Study objective	Target population (demographic data)	No. of patients	HS	Infection site	TDS / Validation	Cal.	Features
					antibiotic susceptibility predictions							
Garcia-Vidal C, 2021 [41]	RCS	Spain	1	Jan 2008-Dec 2017	To assess risk factors for MDR-GNB infections by analyzing a large amount of data and to determine whether ML could be useful in predicting the risk of MDR-GNB infection at the onset of febrile neutropenia	Hematological patients with febrile neutropenia	349 (3,235 episodes of febrile neutropenia)	I	Bloodstream	Yes / Internal	No	28/43
Goodman K-E, 2022 [37]	RCS	USA	1	July 2017-Dec 2019	To understand which patient and treatment characteristics are associated with either a higher or lower likelihood of intervention in a PAF program and to develop prediction models to identify antimicrobial orders that may be excluded from the review	Patient with antimicrobial orders from University of Maryland Medical Centre	17,503; testing set 3,435	I	Bloodstream, bone/ joint, central nervous system, cardiac/ vascular, gastrointestinal, genitourinary, respiratory	Yes / Internal	No	33
Herman B, 2021 [58]	XS	Indonesia	Multiple	Model Building Jan 2015-Dec 2019; Testing Jan 2020-Oct 2020	To develop and evaluate the performance of the CUHAS-ROBUST, an AI-based tool for screening RR-TB, particularly in resource-limited settings where rapid diagnostic tests are not widely available	Patients with suspected RR-TB	487 for model building + 157 for testing (644 total)	I	Respiratory	Yes / Internal	No	19
Huang T-S, 2020 [44]	CCS	Taiwan	1	Jan 2016-Oct 2017	Detection of CR *K. pneumoniae* on the basis of MALDI	NA	46 CR *K. pneumoniae*; 49 *K. pneumoniae* susceptible	NA	NA	Yes / Internal	No	NA
İlhanlı N, 2024 [54]	RCS	South Korea	1	Oct 2012-Oct 2022	To predict antibiotic resistance in patients with UTI using ML models and to interpret these models	Patients with UTI, aged 19 to 100 years	3,865 (cephalosporin 3,535, TZP 737, TMP-SMX 708, fluoroquinolone 1,582, CAR 1,365)	I / O	UTI	Yes / Internal	No	71
Jiménez F, 2020 [42]	ERS	Spain	1	Jan 2009-Jan 2018	To develop and evaluate ML models, particularly multivariate time series forecasting with feature selection, to predict antibiotic resistance outbreaks in hospitals using data from a single center	Patients who were at risk of or had developed antibiotic-resistant infections, particularly within the hospital setting	NA	I	NA	Yes / Internal	No	Initially: 508; SVM: 13, DT: 7, RF: 5
Lee A-L-H, 2021 [59]	RCS	Hong Kong	3	Jan 2015-Dec 2019	To predict ESBL production in community-onset Enterobacteriaceae bacteremia	Patients with community-onset bacteremia	5,625; split into a training set and a test set in a 4:1 ratio	I	NA	Yes / Internal	No	133
Liang Q-Q, 2022 [47]	RCS	China	2	2015–2019	To develop a model using ML to predict CR-GNB carriage within one week in ICU patients and to validate this model prospectively	ICU patients with potential CR-GNB carriage	10,247; 1,385 with positive CR-GNB cultures, 1,535 with negative CR-GNB cultures	I	Respiratory	Yes / External	No	16
Liang Q-Q, 2024 [48]	RCS	China	1	Jan 2015-Dec 2021	To predict whether the pathogen causing bloodstream infections in the ICU is CR-GNB using ML algorithms	Critically ill patients admitted to the ICU with suspected bloodstream infections	952 with positive blood cultures (CR-GNB 418, non-CR-GNB 534), 1,422 with negative blood cultures; patients were divided into training (70 %), validation (15 %), and test (15 %) sets	I	Bloodstream	Yes / Internal	NA	NA

(*continued on next page*)

Table 2 (*continued*)

Author, year (Ref)	Study type	Country	No. of centers	Data time frame	Study objective	Target population (demographic data)	No. of patients	HS	Infection site	TDS / Validation	Cal.	Features
Martínez-Agüero S, 2019 [43]	RCS	Spain	1	2003–2015	To develop ML models to predict antimicrobial resistance in *P. aeruginosa*; the study aimed to provide early identification of resistant bacteria to improve patient outcomes and reduce the spread of resistant infections	ICU patients	Total patients 2,630: AMG: 2,177; CAR: 1,458; Fourth-generation cephalosporins: 1,582; broad-spectrum antibiotics: 2,309; POLY: 570; QUI: 1,952	I	Bloodstream	Yes / NA	NA	20–30/ 78–127
Nigo M, 2024 [38]	RCS	USA	2	Jan 2018-Apr 2021	To make an accurate risk stratification of MRSA	Patients from Memorial Hermann Hospital System, Houston; validation with database MIMIC-IV	Memorial Hermann Hospital System: 26,233; for validation database MIMIC-IV: 152,979	I	NA	Yes / External	No	NA
Oonsivilai M, 2018 [55]	RCS	Cambodia	1	Feb 2013-Jan 2016	To predict Gram stains and whether bacterial pathogens could be treated with standard empirical antibiotic regimens	Children with at least one positive blood culture from Angkor Hospital for Children	AMP + GEN: 243; ceftriaxone: 68	I	Bloodstream	Yes / Internal	Yes	35
Shang J-S, 2000 [39]	RCS	USA	5	Mar 1996-Mar 1997	To investigate the potential of using NN and LR approach in diagnosing MRSA	*S. aureus*-infected patients in our study were obtained from five medical facilities in the Pittsburgh area	504	I	NA	Yes / Internal	No	8
Tacconelli E, 2020 [60]	PCS	Italy, Serbia, and Romania	3	Sep 2010-Jun 2013	To measure the impact of antibiotic exposure on the acquisition of colonization with ESBL-GNB, accounting for individual- and group-level confounding using ML methods	Hospitalized patients in medical and surgical wards	10,034 (28,322 rectal swabs)	I	Gastrointestinal	Yes / NA	NA	39
Tsai W-C, 2023 [45]	RCS	Taiwan	1	Jan 2017-Dec 2020	To evaluate the ability of the RF algorithm to predict bacteremia in adult febrile patients in the ED	Adult febrile patients (aged $\geq$ 20 years) with at least one blood culture taken at the ED	5,647; divided into derivation (3,369) and validation (2,278)	E	NA	Yes / Internal	NA	21
Tsurumi A, 2023 [40]	RCS	USA	7	2003–2009	To facilitate precision medicine plans in the pediatric burn-care setting and to create innovative predictive tools for early prophylactic and therapeutic interventions in thermally-injured children	Children with burns	82	E	Bloodstream	No / Internal	No	NA
Wang C-X, 2022 [46]	RCS	Taiwan	1	Jun 2013-Feb 2018	To develop and validate ML models for rapid detection of CIRKP using mass spectrometry data combined with patient demographics	Patients infected with *K. pneumoniae*	16,697 samples initially, with 15,782 samples selected after quality control (11,354 CISKP and 4,428 CIRKP)	I	Bloodstream, wound, respiratory, UTI	Yes / Internal	No	102/480
Wong J-G, 2020 [56]	PCS	Singapore	1	Jun 2016-Nov 2018	To develop prediction models based on local clinical and laboratory data to guide antibiotic prescribing for adult patients with uncomplicated upper respiratory tract infections	Patients with uncomplicated URTI at the ED at Tan Tock Seng Hospital	715; using 70 % of the participants as training set	E	Respiratory	Yes / Internal	No	50

Abbreviations

AI: artificial intelligence, AMG: aminoglycosides, AMP: ampicillin, APSS: antimicrobial prescription surveillance system, ASP: antibiotic stewardship program, Cal.: calibration, CAR: carbapenems, CCS: case-control study, CIRKP: ciprofloxacin-resistant *Klebsiella pneumoniae*, CISKP: ciprofloxacin-susceptible *Klebsiella pneumoniae*, CDSS: clinical decision support system, CR: carbapenem-resistant, CRE: carbapenem-resistant Enterobacteriaceae, CT: Computed Tomography, CUHAS-ROBUST: Chulalongkorn-Hasanuddin Rifampicin Resistant Tuberculosis Screening Tool, DT: Decision Tree, E: Emergency, EAT: empirical antibiotic therapy, ED: emergency department, EHR: electronic health records, ERS: experimental research study, ESBL-PE: extended-spectrum beta-lactamase-producing Enterobacteriaceae, UAE: United Arab Emirates, GEN: gentamicin, GNB: Gram-

studies included. The most common were laboratory/microbiological data (96.3 %, n = 26) [34,36–60], followed by clinical data (92.6 %, n = 25) [34–43,45–52,54–60]. Only 11.1 % (n = 3) of the studies included pediatric populations [40,51,55].

3.4. ML algorithms and performance

Table 3 summarizes algorithms used in the studies included and also provides a first comparison between ML algorithms and logistic regression regarding AUC outcome. The group of Decision Tree algorithms was the most applied, with 29 uses. The most used algorithm, both of the group of Decision Tree and also in general, is RF (n = 19) [35–37,41–48,51–55,57,58,60]. Other algorithms of this group used are DT (n = 7) [36,43,47,48,55,58,60], J48(C4.5) (n = 2) [53,57], and CART (n = 1) [56]. Boosted models were applied 13 times, with XGB used in 9 cases [36,41,45–48,51,52,58], along with other Boosting algorithms (Boosted DT, n = 2 [50,55]; GBM, n = 2 [34,41]). SVM models were used 12 times (most frequently SVM, n = 6 [42,44,46,48,51,52]; SMO, n = 2 [53,57]; and Linear SVM, n = 1 [55]). Neural Networks, including both general NN (Artificial NN, n = 2 [42,58]; NN, n = 2 [39,59]) and more advanced variants such as Recurrent NNs (Recurrent NN, n = 1 [38]), were also employed, totaling five instances.

In addition, the table provides a full quantitative summary of the predictive performance of various ML models and logistic regression, measured by the AUC. Regarding the ML models, the AUC ranges from 54.0 [55] (DT model) to 97.8 [40], while for logistic regression, it spans from 48.0 [51] to 91.1 [46]. It was evaluated in three main contexts: predicting antibiotic resistance using ML models, assessing resistance to specific antibiotics, and predicting pathogen-specific resistance to various antibiotics.

For ML models, accuracy ranged from 51.0 [49] to 94.0 [44] (the latter achieved by an RF model) and for logistic regression from 59.5 [60] to 91.0 [44] (**Table S3**). Sensitivity values varied from 25.0 [58] to 98.7 [34] for ML models, and for logistic regression from 21.2 [58] to 98.4 [46] (**Table S4**). For ML models, specificity ranged from 28.0 [35] to 100.0, with RF [41,44] and Naïve Bayes [44] reaching the upper limit, and for logistic regression from 30.0 [59] to 100.0 [39] (**Table S5**). For ML models, the positive predictive value ranged from 25.0 [45] to 100.0 [41] and the negative predictive value spanned from 32.4 [53] to 99.0 [57], and for logistic regression the positive predictive value varies from 27.6 [40] to 100.0 [39] and the negative predictive value from 43.2 [36] to 94.5 [39] (**Table S6**).

3.5. Meta-analysis

For the AUC, 11 studies and 68 ML and traditional models were analyzed. Using the fixed effects model (FEM), the pooled standardized mean difference measured as Cohen's d was 2.16 (95 % CI: 2.14–2.18, $p < 0.001$) based on 871,496 patients (Fig. 4**a**) with high statistical heterogeneity ($I^2 = 99.96$, $p \leq 0.001$). No evidence of publication bias was found in this case, as indicated by the funnel plot and Egger's test (intercept −15.45, $p = 0.211$) (Fig. 4**b**). The random effect model (REM) revealed an ES of 0.67 ([95 % CI: −0.33; 1.67]; $p = 0.186$). After applying the trim and fill method, the estimated ES did not materially change.

Regarding accuracy, we evaluated three studies comprising 18 ML and traditional models. The FEM showed a pooled standardized mean difference measured as Cohen's d of 0.26 (95 % CI: 0.21–0.30, $p < 0.001$); again, with high heterogeneity ($I^2 = 99.67$ %) and a total sample of 8,744 patients. The REM yielded an ES of 0.59 (95 % CI: −0.25–1.42, $p = 0.168$). No evidence of publication bias was found, as indicated by the funnel plot and Egger's test (intercept: 7.10, $p = 0.620$). Results are illustrated in Fig. 5 (a: Forest plot, b: Funnel plot).

For sensitivity, seven studies with 34 ML and traditional models were included for a total of 24,002 patients. The FEM showed an ES of 1.90 (95 % CI: 1.86–1.94, $p < 0.001$), alongside substantial heterogeneity (I^2

negative bacteria, HS: hospital setting, ICU: intensive care unit, I: inpatient, IMD: invasive mold disease, LR: logistic regression, MALDI: Matrix-Assisted Laser Desorption Ionization, MDRO: multidrug-resistant organism, MDR-GNB: multidrug-resistant Gram-negative bacteria, MIMIC: Medical Information Mart for Intensive Care, ML: machine learning, MRSA: methicillin-resistant *Staphylococcus aureus*, NA: not applicable, NN: neural network, O: outpatient, PAF: post-prescription antimicrobial review and feedback, PCS: prospective cohort study, POLY: polymyxins, QUI: quinolones, RCS: retrospective cohort study, RF: random forest, RR-TB: rifampicin-resistant tuberculosis, SMX: sulfamethoxazole, SVM: Support Vector Machine, TDS: training data set, TMP: trimethoprim, TZP: piperacillin/tazobactam, UMC: University Medical Center, URTI: upper respiratory tract infection, UTI: urinary tract infection, VRE: vancomycin-resistant enterococci, XS: cross-sectional study.

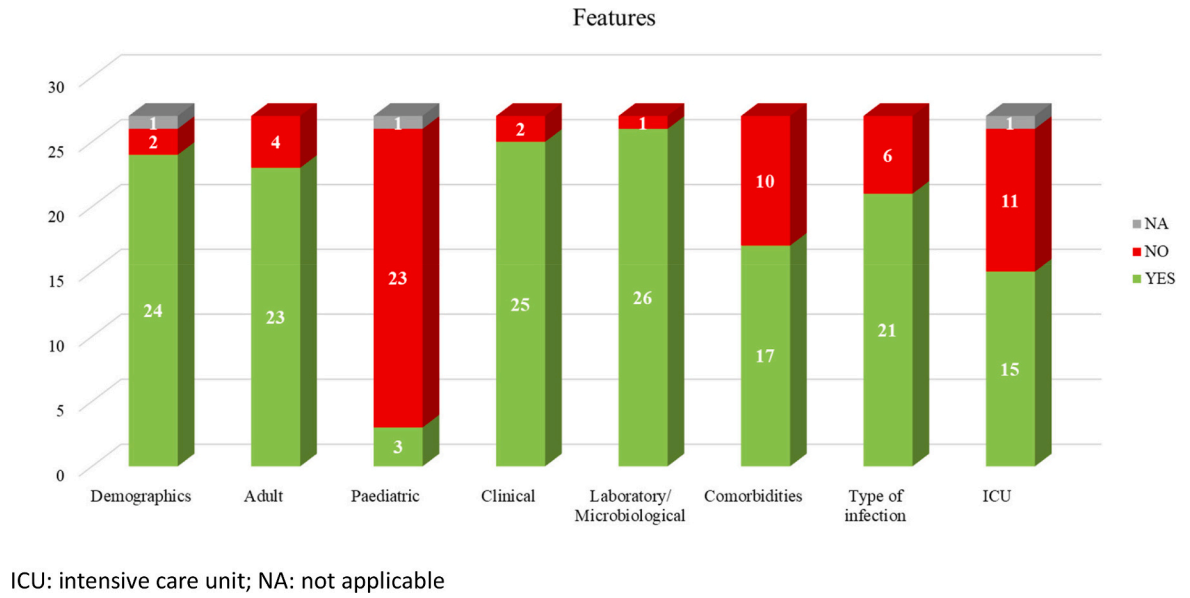


Fig. 3. Distribution of the main features included in the machine learning models. ICU: intensive care unit; NA: not applicable.

= 99.92 %, $p < 0.001$). The REM, however, indicated a comparable ES of 1.93 (95 % CI: 0.48–3.39, $p = 0.009$). No publication bias was detected, as confirmed by Egger's test (intercept: −1.60, $p = 0.942$). Corresponding plots are shown in Fig. 6**a-d**.

When assessing specificity, seven studies encompassing 30 ML and traditional models were evaluated, including 24,858 patients. The FEM provided an ES of −1.39 (95 % CI: −1.42 – −1.36, $p < 0.001$), although heterogeneity remained high ($I^2 = 100$ %, $p < 0.001$) (Fig. 7**a and 7b**). The REM showed a significantly lower ES of −0.22 (95 % CI: −1.31–0.86, $p = 0.688$), based on a total of 24,858 participants.

For the positive predictive value, four studies with 16 ML and traditional models were analyzed. The FEM yielded an ES of −0.61 (95 % CI: −0.64 – −0.58, $p < 0.001$), with high heterogeneity ($I^2 = 99.10$ %, $p < 0.001$) based on 20,006 participants. The REM showed a higher ES of −0.33 (95 % CI: −0.69–0.02, $p = 0.063$). No publication bias was detected (Egger's test: intercept 6.63, $p = 0.400$). These results are shown in Fig. 8**a and 8b**.

Lastly, for the negative predictive value, four studies with 16 ML and traditional models were included. The FEM reported an ES of 1.01 (95 % CI: 0.97–1.04, $p < 0.001$), with high heterogeneity ($I^2 = 99.77$ %, $p < 0.001$), and a sample size of 20,006 patients. The REM demonstrated a higher ES of 1.66 (95 % CI: 0.86–2.46, $p < 0.001$). No publication bias was identified (Egger's test: intercept 18.30, $p = 0.173$). Results are shown in Fig. 9 **a-d**.

3.6. Sensitivity analysis

Sensitivity analysis by study design was not conducted, as all included studies were cross-sectional. Instead, we performed sensitivity analysis based on the type of ML models used. For AUC, when restricted to studies using decision tree models (CART, RF, J48 [C4.5], DT), six studies and 22 ML and traditional models were analyzed, covering a total of 41,904 patients. The FEM demonstrated an ES of 3.88 (95 % CI: 3.84–3.92; $p < 0.001$), although there was considerable heterogeneity ($I^2 = 99.97$ %, $p < 0.001$). The REM showed a lower ES of −0.50 (95 % CI: −3.45 – −2.46, $p = 0.742$) (Fig. 10a and **Fig. S1 a-d**). Conversely, studies employing boosted models (XGB, GBM, CatBoost, Boosted LR, Boosted DT), comprising three studies and eight ML and traditional models with a sample of 21,472 patients, showed an ES of 1.12 (95 % CI: 1.09–1.14, $p < 0.001$) under the FEM. The REM showed an ES of 1.27 (95 % CI: 0.48–2.05, $p = 0.002$) (**Fig. S2 a-d**).

For sensitivity, studies using decision tree models (four studies, 12 ML and traditional models) with a sample of 4,170 patients, showed an ES of 1.13 (95 % CI: 1.06–1.21, $p < 0.001$) in the FEM and an ES of 0.92 (95 % CI: −0.73–2.58, $p = 0.275$) in the REM (**Fig. S3 a-d**). In contrast, three studies using general neural networks (Artificial NN, NN, MLP), with six ML and traditional models and a sample size of 12,572 patients, showed an ES of 7.37 (95 % CI: 7.27–7.47, $p < 0.001$) for the FEM. The REM indicated an improved ES of 6.93 (95 % CI: 2.93–10.92, $p < 0.001$) (Fig. 10b and **Fig. S4 a-d**).

For specificity, four studies using decision tree models (RF, CART, DT, J48) were included, covering six models and a sample of 5,654 patients. The FEM showed an ES of −0.71 (95 % CI: −0.77 – −0.65, $p < 0.001$); the REMs showed a higher ES of −0.52 (95 % CI: −2.05–1.01, $p = 0.505$) (**Fig. S5 a-d**).

To further investigate the various applications of ML, a secondary sensitivity analysis was performed. Only the group with the objective of optimization of prescriptions and clinical decision support met the minimum number of studies required for this analysis. Four studies were included, comprising a total sample of 37,248 patients. The FEM yielded an ES of 6.24 (95 % CI: 6.19–6.29, $p < 0.001$), while the REM demonstrated a lower ES of 2.98 (95 % CI: −0.73–6.69, $p = 0.116$) (**Fig. S6 a-d**).

3.7. Quality assessment of included studies: QUADAS-AI

A thorough analysis of the risk of bias assessment and concerns related to applicability was conducted for each study, with the results summarized in Fig. 11a and Fig. 11b. As for the risk of bias, seven studies (25 %) [36,42–44,50,56,57] had a high risk of bias in patient selection mostly due to the lack of a clear rationale for its sample size and the unspecified data source. Over half of the studies (54 %) exhibited a high risk of bias in the index test domain, primarily due to the absence of external validation or testing. In the reference standard domain, which evaluates the reliability and appropriate application of the gold standard for diagnosis or outcome measurement, 11 % of articles [40,44,56] were at a high risk of bias. Finally, the risk of bias in the flow and timing domain was high in 21 % of studies [39,42,44,50,51,53]. As for applicability concerns, in the patient selection domain, concerns about applicability were low in only 25 % of studies [38–40,51,53,57,59]. In the index test domain, concerns about applicability were high in 50 % of studies due to the lack of detail on the construct or architecture of the algorithm. Finally, in the reference standard domain, concerns about applicability were high in 11 % of studies [40,42,56]. The primary

Table 3
Summary of machine learning algorithms and their predictive performance (corresponding area under the curve values with 95% confidence intervals, where available) compared with logistic regression in antimicrobial stewardship applications.

Author, year	Machine learning algorithms	% AUC – Machine learning algorithms	% AUC – Logistic regression
Ananda-Rajah M−R, 2017 [57]	Naive Bayes, RF, J48(C4.5), SMO	Baseline Text Classifier: 73.9 [95 % CI 67.1–80.6]; Naive Bayes: 92.8 [95 % CI 88.0–97.5]; RF: 94.1 [95 % CI 89.8–98.3]; SMO: 92.4 [95 % CI 87.5–97.3]; J48 (C4.5): 87.2 [95 % CI 80.6–93.7]	LR: 91.1 [95 % CI 85.7–96.5]
Beaudoin M, 2016 [49]	Combined system, Learning modules	NA	NA
Bhavani S, 2020 [34]	GBM	Bacteremia Prediction: 78.0 [95 % CI 77.0–78.0]	LR: 73.0 [95 % CI 72.0–74.0]
Brown D-G, 2023 [35]	RF	10-Features RF Model: 67.0 [95 % CI 66.0–68.0]; 4-Features RF Model: 63.0 [95 % CI 62.0–64.0]	10-Feature LR Model: 70.0 [95 % CI 69.0–71.0]; 4-Feature LR Model: 68.0 [95 % CI 67.0–69.0]
Bystritsky R-J, 2020 [50]	Boosted DT	75.0 [95 % CI 72.0–79.0]	LR 73.0 [95 % CI 69.0–77.0]
Çağlayan Ç, 2022 [36]	RF, XGB, DT	RF VRE 77.0; XGB VRE 77.0; RF MRSA 70.0; XGB MRSA 66.0; RF CRE 72.0; XGB CRE 72.0; RF MDRO 89.0; XGB MDRO 87.0; DT MDRO 81.0	LR VRE: 80.0; LR MRSA: 66.0; LR CRE: 78.0; LR MDRO: 70.0
de Vries S, 2022 [51]	SVM, XGB, RF, k-NN	SVM: 78.2 [sd ± 0.9]; XGB: 80.1 [sd ± 1.1]; RF 80.0 [sd ± 1.0]; k-NN: 78.6 [sd ± 1.0]	LR: 77.4 [sd ± 1.1]
Eickelberg G, 2020 [52]	RF, XGB, MLP, SVM, k-NN, VotingEnsemble	72 h: RF: 79.3; XGB: 79.5; MLP: 77.9; SVM: 77.8; k-NN: 73.4; VotingEnsemble: 79.3 48 h: RF: 78.8; XGB: 76.9; MLP: 77.1; SVM: 77.3; k-NN: 73.3; VotingEnsemble: 78.824 h: RF: 77.4; XGB: 77.6; MLP: 76.4; SVM: 76.3; k-NN: 71.4; VotingEnsemble: 77.0	LR 72 h: 87.1; LR 48 h: 77.4; LR 24 h: 76.4
Feretzakis G, 2020 [53]	SVM C-support Vector; SMO; k-NN; J48(C4.5); RF; RIPPER; MLP	Library SVM C-Support Vector 66.0; SMO 65.9; 1-Nearest Neighbors 68.2; 5-Nearest Neighbors 71.1; J48(C4.5) 72.4; RF 70.3; RIPPER 69.9; MLP 72.6	LibraryLINEAR 56.8
Garcia-Vidal C, 2021 [41]	RF, GBM, XGB	RF: 78.9; GBM: 78.7; XGB: 79.4	Generalized Linear Model: 78.3
Goodman K-E, 2022 [37]	RF	RF 76.0 [95 % CI 75.0–77.0]	LR 70.0 [95 % CI 68.0–72.0]
Herman B, 2021 [58]	Artificial NN, DT, RF, XGB	Artificial NN 96.0 in training data	NA
Huang T-S, 2020 [44]	RF, Naïve Bayes, k-NN, SVM	NA	NA
İlhanlı N, 2024 [54]	RF	RF − Training set: cephalosporin: 77.7 [95 % CI 77.5–77.9]; TZP: 86.4 [95 % CI 86.2–86.7]; CAR: 87.7 [95 % CI 87.4–88.0]; TMP-SMX: 88.1 [95 % CI 87.9–88.2]; fluoroquinolone: 88.4 [95 % CI 88.4–88.5]RF − Test set: cephalosporin: 63.8 [95 % CI 63.5–64.2]; TZP: 63.0 [95 % CI 62.6–63.4]; CAR: 66.5 [95 % CI 65.9–67.1]; TMP-SMX: 67.0 [95 % CI 66.6–67.3]; fluoroquinolone: 72.1 [95 % CI 71.8–72.4]	LRTraining Set: cephalosporin: 71.8; TZP: 61.5; CAR: 63.8; TMP-SMX: 76.6; fluoroquinolone: 79.1Test Set: cephalosporin: 61.5; TZP: 53.8; CAR: 56.5; TMP-SMX: 63.3; fluoroquinolone: 70.4
Jiménez F, 2020 [42]	Gaussian Processes, Artificial NN[q], RF[s], SVM[w]	SVM: 86.6; RF: 82.0; Artificial NN: 80.0; Gaussian Processes: 81.0	LR: 76.0
Lee A-L-H, 2021 [59]	NN	76.1 [95 % CI 72.5–79.7]	LR: 66.7 [95 % CI 62.7–70.7]
Liang Q-Q, 2022 [47]	DT, RF, XGB	RF: validation set 91.0, test set 90.0; XGB: validation set 91.0, test set 89.0; DT: validation set 90.0, test set 89.0	LR: 81.0 (validation set), 78.0 (test set)
Liang Q-Q, 2024 [48]	RF, SVM, DT, XGB	Bloodstream Infection Model: DT: 77.0, RF: 86.0, SVM: 83.0, XGB: 85.0; CR-GNB Bacteremia Model: DT: 69.0, RF: 87.0, SVM: 88.0, XGB: 80.0	Bloodstream Infection Model: LR[l]: 81.0;CR-GNB Bacteremia Model: LR: 86.0
Martínez-Agüero S, 2019 [43]	k-NN, DT, RF, MLP	NA	NA
Nigo M, 2024 [38]	Recurrent NN	91.1 [95 % CI 90.0–91.6]	LR: 85.7 [95 % CI 84.9–86.5]
Oonsivilai M, 2018 [55]	RF, DT, Boosted DT, Linear SVM, Polynomial SVM, SVM with RBF kernel, k-NN	Resistant to ceftriaxone: RF: 80.0 [95 % CI 66.0–94.0]; Boosted DT: 88.0 [95 % CI 77.0–100.0]; DT: 79.0 [95 % CI 65.0–92.0]; Regularized LR: 75.0 [95 % CI 58.0–91.0]; k-NN: 59.0 [95 % CI 42.0–77.0]; SVM Radial Kernel: 80.0	Resistant to ceftriaxone LR[l]: 81.0 [95 % CI 65.0–96.0]; Resistant to AMP + GEN[h] LR[l]: 48.0 [95 % CI 30.0–65.0]; Gram stain LR[l]: 58.0 [95 % CI 41.0–74.0]

(continued on next page)

Table 3 (*continued*)

Author, year	Machine learning algorithms	% AUC – Machine learning algorithms	% AUC – Logistic regression
		[95 % CI 64.0–96.0]; SVM Linear Kernel: 81.0 [95 % CI 68.0–95.0]; SVM Polynomial Kernel: 83.0 [95 % CI 71.0–95.0] Resistant to AMP + GEN: RF: 74.0 [95 % CI 59.0–89.0]; Boosted DT: 61.0 [95 % CI 43.0–78.0]; DT: 54.0 [95 % CI 46.0–63.0]; Regularized LR: 59.0 [95 % CI 43.0–75.0]; k-NN: 60.0 [95 % CI 43.0–76.0]; SVM Radial Kernel: 56.0 [95 % CI 38.0–73.0]; SVM Linear Kernel: 58.0 [95 % CI 41.0–75.0]; SVM Polynomial Kernel: 63.0 [95 % CI 46.0–80.0]Gram stain: RF: 71.0 [95 % CI 57.0–86.0]; Boosted DT: 62.0 [95 % CI 46.0–79.0]; DT: 63.0 [95 % CI 48.0–77.0]; Regularized LR: 66.0 [95 % CI 51.0–82.0]; k-NN: 61.0 [95 % CI 45.0–77.0]; SVM Radial Kernel: 68.0 [95 % CI 53.0–84.0]; SVM Linear Kernel: 75.0 [95 % CI 60.0–89.0]; SVM Polynomial Kernel: 75.0 [95 % CI 62.0–89.0]	
Shang J-S, 2000 [39]	NN	NN 92.8 [sd ± 1.6]	LR: 86.9 [sd ± 1.9]
Tacconelli E, 2020 [60]	RF, DT	RF: validation set: 91.8, test set: 90.8; XGB: validation set: 91.2, test set: 89.4; DT: validation set: 90.7, test set: 89.8	LR: validation set: 81.7, test set: 78.4
Tsai W-C, 2023 [45]	RF, MLP, XGB, LightGBM	RF: 76.1; MLP: 74.5; XGB: 74.0; LightGBM: 73.1; RF (validation group): 70.9	LR: 75.5
Tsurumi A, 2023 [40]	LASSO	Multibiomarker panel: 93.8 [95 % CI 88.1–98.1]; Multibiomarker panel + TBSA: 96.5 [95 % CI 93.9–99.9]; Multibiomarker panel + TBSA + inhalation injury status: 97.8 [95 % CI 94.1–100.0]	TBSA + inhalation injury status: 76.5 [95 % CI 74.7–78.3]
Wang C-X, 2022 [46]	SVM, XGB, RF	SVM: 89.0; XGB: 89.0; RF: 85.0	LR: 85.0
Wong J-G, 2020 [56]	LASSO, CART	LASSO 70.0 [95 % CI 62.0–77.0]; CART 67.0 [95 % CI 59.0–74.0]	LR 72.0 [95 % CI 65.0–79.0]

Abbreviations

AMP: ampicillin, CAR: carbapenems, CART: classification and regression trees, CI: confidence interval, CR: carbapenem-resistant, CRE: carbapenem-resistant Enterobacteriaceae, DT: decision tree, GBM: Gradient Boosting Machine, GEN: gentamicin, GNB: Gram-negative bacteria, k-NN: k-Nearest Neighbors, LASSO: Least Absolute Shrinkage and Selection Operator, LR: logistic regression, MDRO: multidrug-resistant organism, ML: machine learning, MLP: Multilayer Perceptron, MRSA: methicillin-resistant *Staphylococcus aureus*, NN: Neural Network, RBF: Radial Basis Function, RF: Random Forest, RIPPER: Repeated Incremental Pruning to Produce Error Reduction, SMO: Sequential Minimal Optimization, SMX: sulfamethoxazole, SVM: Support Vector Machine, TBSA: Total Burn Surface Area, TMP: trimethoprim, TZP: piperacillin/tazobactam, VRE: vancomycin-resistant enterococci, XGB: Extreme Gradient Boosting.

factors contributing to lower applicability included using surrogate or proxy measures instead of gold-standard diagnostic tests, sole reliance on internal validation, and patient cohorts that may not accurately represent the general clinical population. Furthermore, problems such as unclear inclusion or exclusion criteria or a patient population that does not reflect real-world clinical settings can introduce bias and further diminish applicability.

4. Discussion

4.1. Interpretation of the main results

This systematic review and meta-analysis synthesizes current evidence on the application of ML models in AMS, offering a comparative assessment with traditional statistical approaches. The findings demonstrate that ML models generally achieve superior performance in key metrics such as sensitivity and predictive accuracy. However, these advantages are accompanied by significant limitations, including high heterogeneity across studies, limited external validation, and challenges related to model interpretability and calibration. Addressing these issues is essential to ensure the safe, effective, and equitable integration of ML into clinical practice.

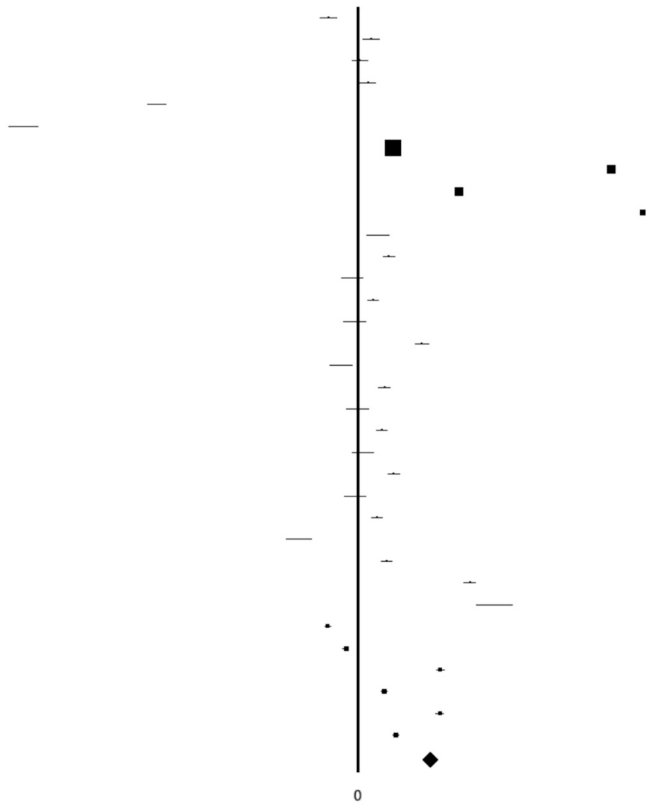
Across multiple studies, ML models achieved higher sensitivity, specificity, and predictive accuracy, with decision tree and boosted methods reaching AUC values over 90 %. However, inconsistencies with FEM or REMs limit generalizability. ML effectiveness depends on data quality, algorithm choice, and clinical context, requiring further validation and standardization to move from experimental success to practical use.

The meta-analytic synthesis revealed significant ESs favoring ML models for predictive accuracy and sensitivity, though high heterogeneity across studies likely reflects variability in design, population, and implementation. These conclusions remain robust due to minimal publication bias.

While traditional models sometimes showed competitive specificity in low-complexity cases, ML models excel in complex scenarios, like predicting multidrug resistance, by integrating high-dimensional datasets. However, the meta-analysis highlighted gaps such as limited external validation and underrepresentation of outpatient and pediatric populations, which must be addressed to improve the generalizability of ML approaches.

a)

	ES	95% CI	Sig.	N
Ananda-Rajah M-R, J48(C4.5) 2017	-0.87	-1.13 , -0.61	0.000	246
Ananda-Rajah M-R, Naive Bayes 2017	0.41	0.15 , 0.66	0.002	246
Ananda-Rajah M-R, RF 2017	0.07	-0.18 , 0.32	0.568	246
Ananda-Rajah M-R, SMO 2017	0.31	0.06 , 0.56	0.016	246
Brown D-G, 10-featuresRF 2023	-6.00	-6.28 , -5.72	0.000	1056
Brown D-G, 4-featuresRF 2023	-10.00	-10.44 , -9.56	0.000	1056
Bystritsky R-J, Boosted DT 2020	1.05	1.02 , 1.08	0.000	19230
Goodman K-E, RF 2022	7.59	7.53 , 7.65	0.000	35006
Lee A-L-H, NN 2021	3.03	2.98 , 3.08	0.000	11250
Nigo M, Recurrent NN 2024	8.54	8.45 , 8.63	0.000	787426
Oonsivilai M, Boosted DT1 2018	0.60	0.26 , 0.95	0.001	136
Oonsivilai M, Boosted DT2 2018	0.94	0.75 , 1.12	0.000	486
Oonsivilai M, DT1 2018	-0.17	-0.51 , 0.17	0.327	136
Oonsivilai M, DT2 2018	0.47	0.29 , 0.65	0.000	486
Oonsivilai M, RF1 2018	-0.08	-0.42 , 0.25	0.625	136
Oonsivilai M, RF2 2018	1.92	1.71 , 2.14	0.000	486
Oonsivilai M, Regularized LR1 2018	-0.49	-0.83 , -0.15	0.005	136
Oonsivilai M, Regularized LR2 2018	0.80	0.62 , 0.99	0.000	486
Oonsivilai M, SVM Linear1 2018	0.00	-0.34 , 0.34	1.000	136
Oonsivilai M, SVM Linear2 2018	0.72	0.54 , 0.91	0.000	486
Oonsivilai M, SVM Polynomial1 2018	0.17	-0.17 , 0.51	0.318	136
Oonsivilai M, SVM Polynomial2 2018	1.09	0.89 , 1.28	0.000	486
Oonsivilai M, SVM Radial1 2018	-0.08	-0.42 , 0.25	0.635	136
Oonsivilai M, SVM Radial2 2018	0.58	0.39 , 0.76	0.000	486
Oonsivilai M, k-NN1 2018	-1.76	-2.15 , -1.36	0.000	136
Oonsivilai M, k-NN2 2018	0.87	0.69 , 1.06	0.000	486
Shang J-S, NN 2000	3.36	3.17 , 3.55	0.000	1008
Tsurumi A, LASSO 2023	4.09	3.56 , 4.63	0.000	164
Wong J-G, CART 2020	-0.89	-1.00 , -0.78	0.000	1430
Wong J-G, LASSO 2020	-0.36	-0.46 , -0.25	0.000	1430
de Vries S, RF 2022	2.47	2.34 , 2.60	0.000	1620
de Vries S, SVM 2022	0.80	0.69 , 0.90	0.000	1620
de Vries S, XGB 2022	2.45	2.33 , 2.58	0.000	1620
de Vries S, k-NN 2022	1.14	1.04 , 1.25	0.000	1620
Overall (fixed-effect model)	2.16	2.14 , 2.18	0.000	871496



b)

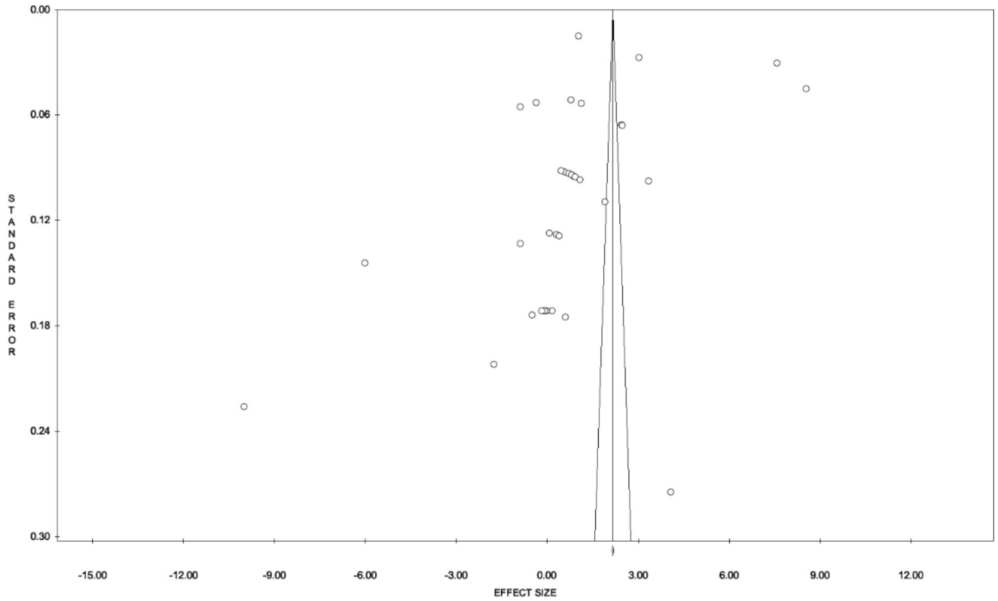


Fig. 4. a) forest plot and b) funnel plot of the fixed effect model assessing the auc, with cohen's d as the effect size measure.

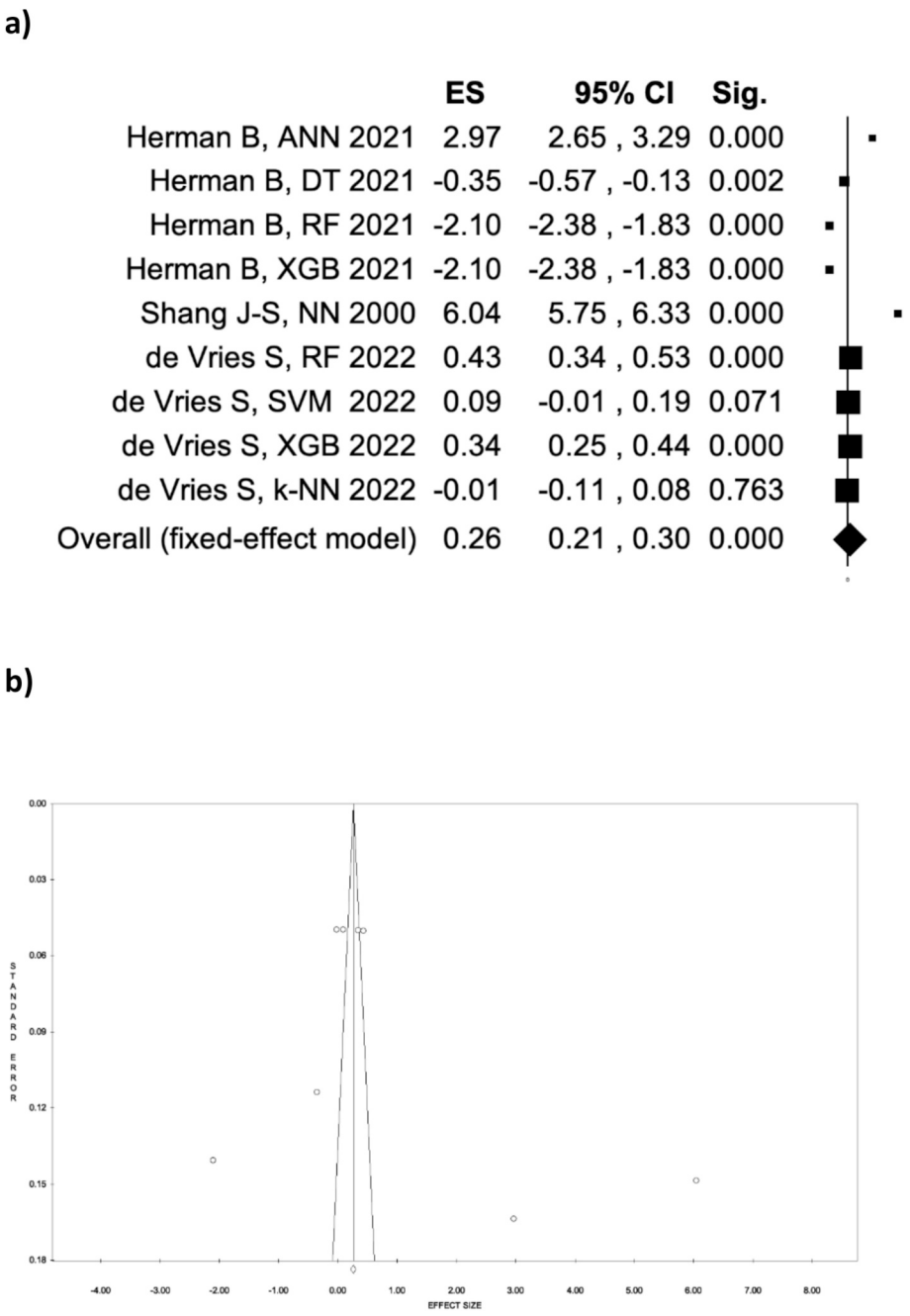


Fig. 5. a) forest plot and b) funnel plot of the fixed effect model assessing the accuracy, with cohen's d as theeffect size (ES) measure.

4.2. Comparison with existing literature

The findings of this systematic review and meta-analysis align with and extend the existing body of literature on the use of ML in AMS. Previous studies have highlighted the promise of ML models in predicting AMR and optimizing prescribing practices, particularly in complex clinical scenarios [61]. For instance, several reviews have reported the superiority of ML algorithms, such as random forests and gradient boosting methods, over traditional logistic regression in terms of predictive accuracy and adaptability to diverse datasets [62]. These observations are consistent with our findings, where ML models demonstrated higher sensitivity and specificity across various study settings.

Our results offer a detailed synthesis of ESs for performance metrics like AUC, sensitivity, and specificity, which were inconsistently reported in earlier literature. Unlike prior single-center or infection-specific analyses, our review highlights the generalizability of ML approaches across infections and populations, though gaps remain for outpatient and pediatric data. Contrasting with some earlier reviews that questioned the clinical applicability of ML models due to concerns about interpretability [63], this analysis emphasizes their practical utility, particularly in high-risk settings such as intensive care units and bloodstream infections. This finding aligns with the growing interest in integrating explainable AI frameworks, which were utilized in a subset of studies included in our review [34,50,53]. Nonetheless, gaps in calibration and external validation remain consistent with previous critiques, highlighting ongoing challenges in ensuring the robustness and reproducibility of ML-driven AMS tools.

Overall, this review substantiates and expands on the existing evidence base, providing a comprehensive assessment of ML's capabilities

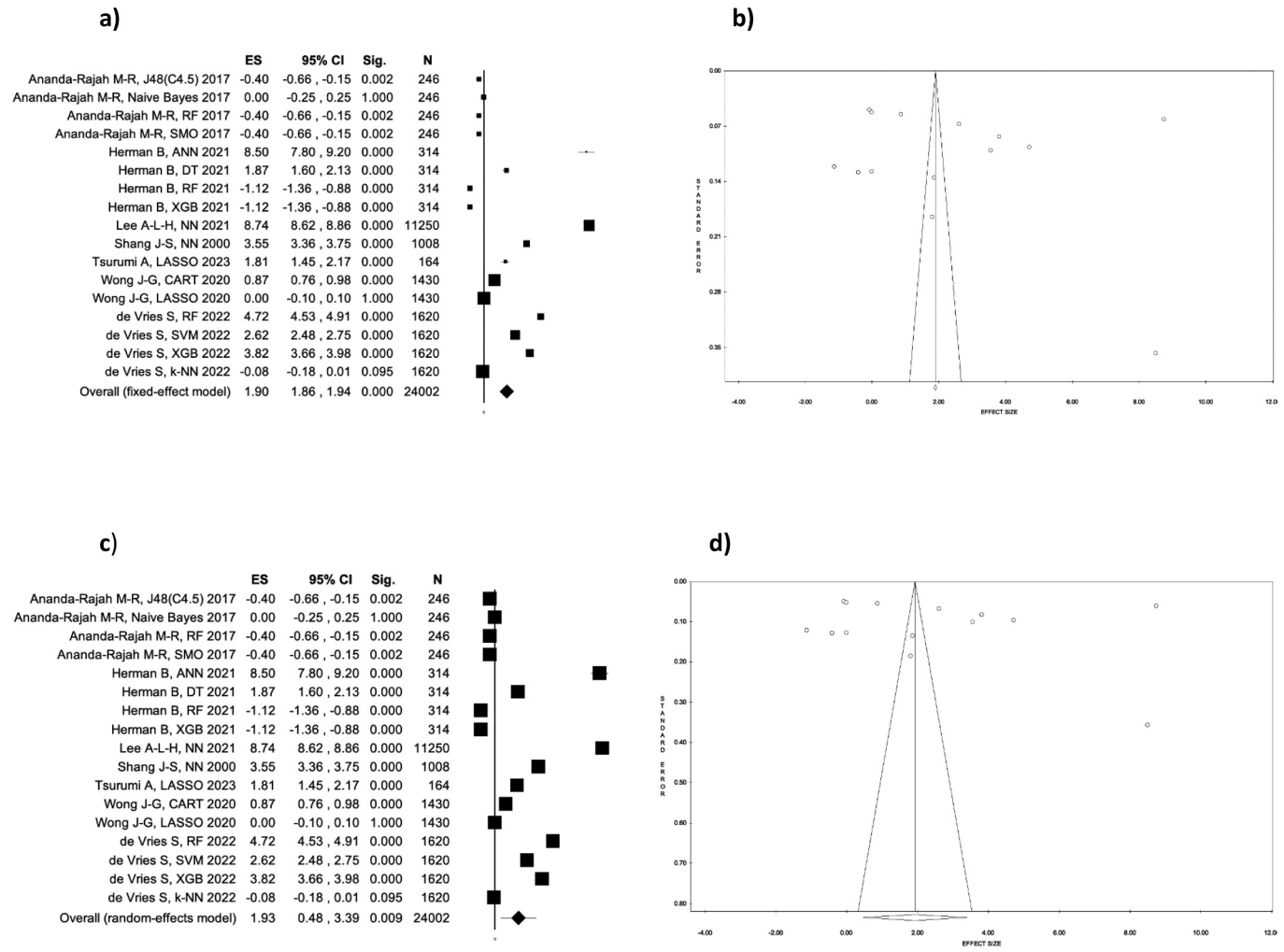


Fig. 6. a) forest plot and b) funnel plot of the fixed effect model and c) forest plot and d) funnel plot of the random effect model assessing the sensitivity, with cohen's d as theeffect size (ES) measure.

while identifying critical areas for future exploration.

4.3. Future research directions

This review highlights several areas where future research is essential to maximize the potential of ML models in AMS. First, there is a pressing need for more studies focusing on external validation to ensure the generalizability and robustness of ML models across diverse clinical settings, particularly in low-resource environments [64]. Current research predominantly relies on internal validation, limiting the applicability of findings beyond single-center studies.

Second, greater emphasis should be placed on integrating ML models into outpatient care and pediatric populations, which are underrepresented in the existing literature. Additionally, future research should explore combining ML algorithms with real-time clinical decision-support systems to streamline implementation and foster adoption by healthcare providers [65].

Third, addressing concerns about the interpretability and transparency of ML models is critical [63]. Development and evaluation of explainable AI frameworks could enhance clinicians' trust and understanding of these tools, promoting their integration into routine clinical workflows. This involves not only creating more interpretable models but also standardizing reporting practices to ensure reproducibility and clarity in ML research.

Future studies should also assess the impact of ML-driven AMS interventions on clinical outcomes [66], such as reducing AMR, improving patient safety, and optimizing healthcare resource utilization. Large-scale, multicenter trials are required to establish the clinical efficacy and cost-effectiveness of these models.

In summary, advancing the field of ML in AMS requires a multifaceted approach that emphasizes external validation, inclusivity of diverse populations, interpretability, and clinical outcomes. Addressing these gaps will pave the way for robust solutions to combat AMR globally.

4.4. Implications for public health policies and prevention

While the findings of this review highlight the transformative potential of ML models in AMS, a critical and cautious approach is warranted, when considering their immediate and complete replacement of traditional systems. Although ML models demonstrate superior predictive accuracy and the ability to process high-dimensional data in real time, their adoption presents several challenges.

First, the reliance on ML models can amplify existing biases if the training data are unrepresentative or flawed. As Yoon *et al.* [63] emphasize, the opacity of "black box" algorithms can obscure errors or reinforce systemic inequities, particularly affecting marginalized populations. These "black-box" models generate predictions through complex, non-transparent processes, making it difficult for clinicians to understand the rationale behind individual outputs. This lack of transparency can reduce trust, hinder clinical acceptance, and complicate accountability in decision making, particularly in high-stakes settings

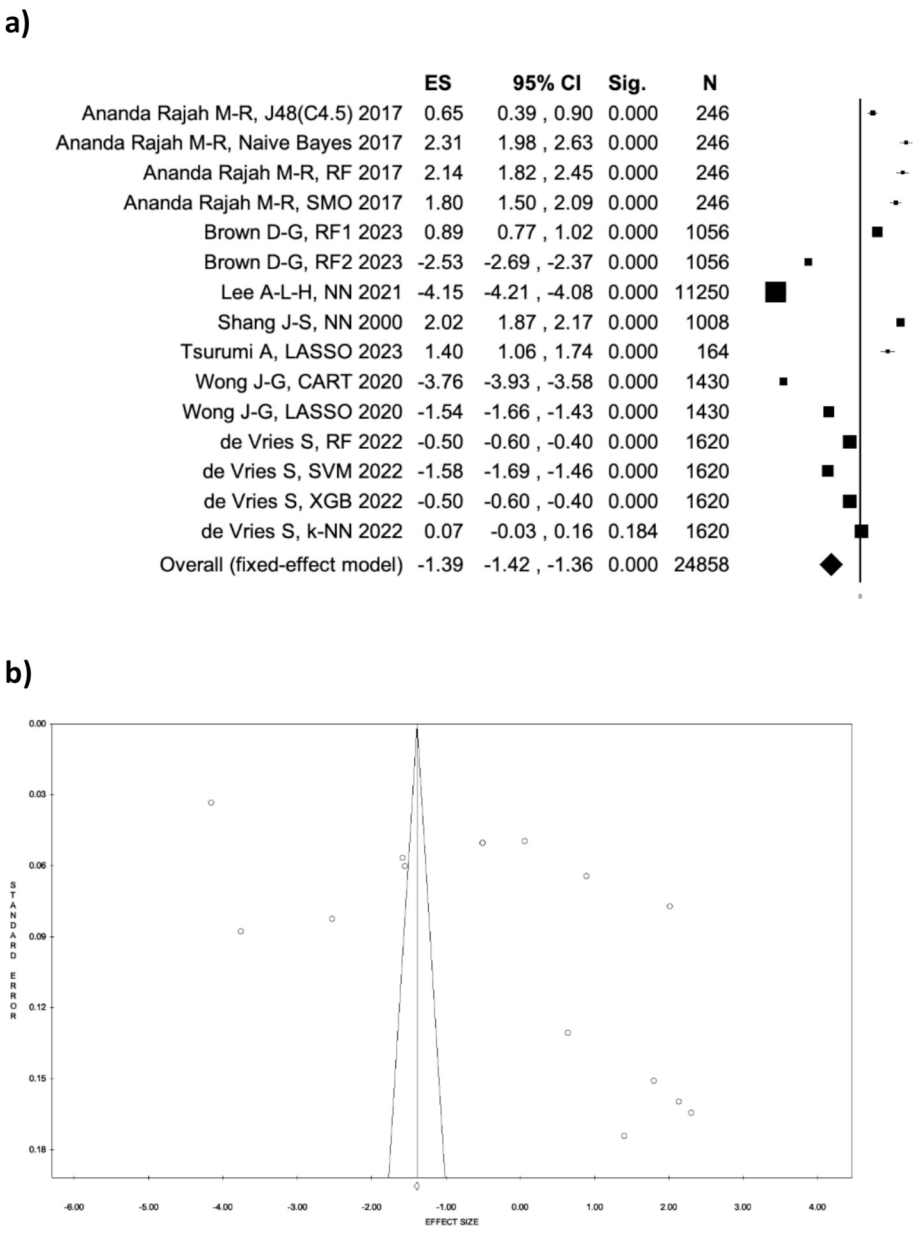


Fig. 7. a) forest plot and b) funnel plot of the fixed effect model assessing the specificity, with cohen's d as theeffect size (ES) measure.

such as antimicrobial prescribing. Furthermore, few authors have discussed the potential of over-fitting when applying these black-box models, which is one of the major concerns when applying these methods [67]. In addition to interpretability, calibration is a crucial but underreported aspect of model performance. Calibration assesses how well predicted probabilities align with actual outcomes, ensuring that risk estimates are reliable for clinical use. Poorly calibrated models may lead to inappropriate antimicrobial decisions, either underestimating or overestimating the risk of resistance, with direct consequences on patient care. Notably, only a minority of studies included in this review reported calibration metrics, highlighting a significant gap in current research.

Future studies should incorporate explainability methods – such as SHAP or LIME – to provide insights into model predictions, alongside routine calibration assessments using tools like calibration plots and Brier scores. Improving both interpretability and calibration is essential to support the safe, trustworthy, and effective integration of ML into AMS programs.

This underscores the need for interpretability and rigorous auditing of ML tools before they are widely deployed, to ensure equitable health outcomes. Interpretability not only fosters transparency but also allows stakeholders to identify and mitigate biases that could exacerbate health inequities. This is especially critical for ensuring that ML applications do not inadvertently perpetuate racial or socioeconomic disparities. Policies should thus ensure that ML systems are interpretable and subject to rigorous auditing for fairness and equity.

Second, the interpretability of ML models remains a barrier to clinical trust and usability [63]. Without a clear understanding of the rationale behind ML-driven recommendations, clinicians may hesitate to fully rely on these systems. This lack of transparency can undermine the integration of ML into AMS programs and risks shifting accountability ambiguously between human operators and automated systems, as noted by Yoon *et al.* [63].

Furthermore, the infrastructure required for ML integration – such as data standardization, computational resources, and clinician training – is substantial, particularly in resource-limited settings [68]. Traditional systems, while less precise, are well-established, familiar to healthcare professionals, and operational without the technological

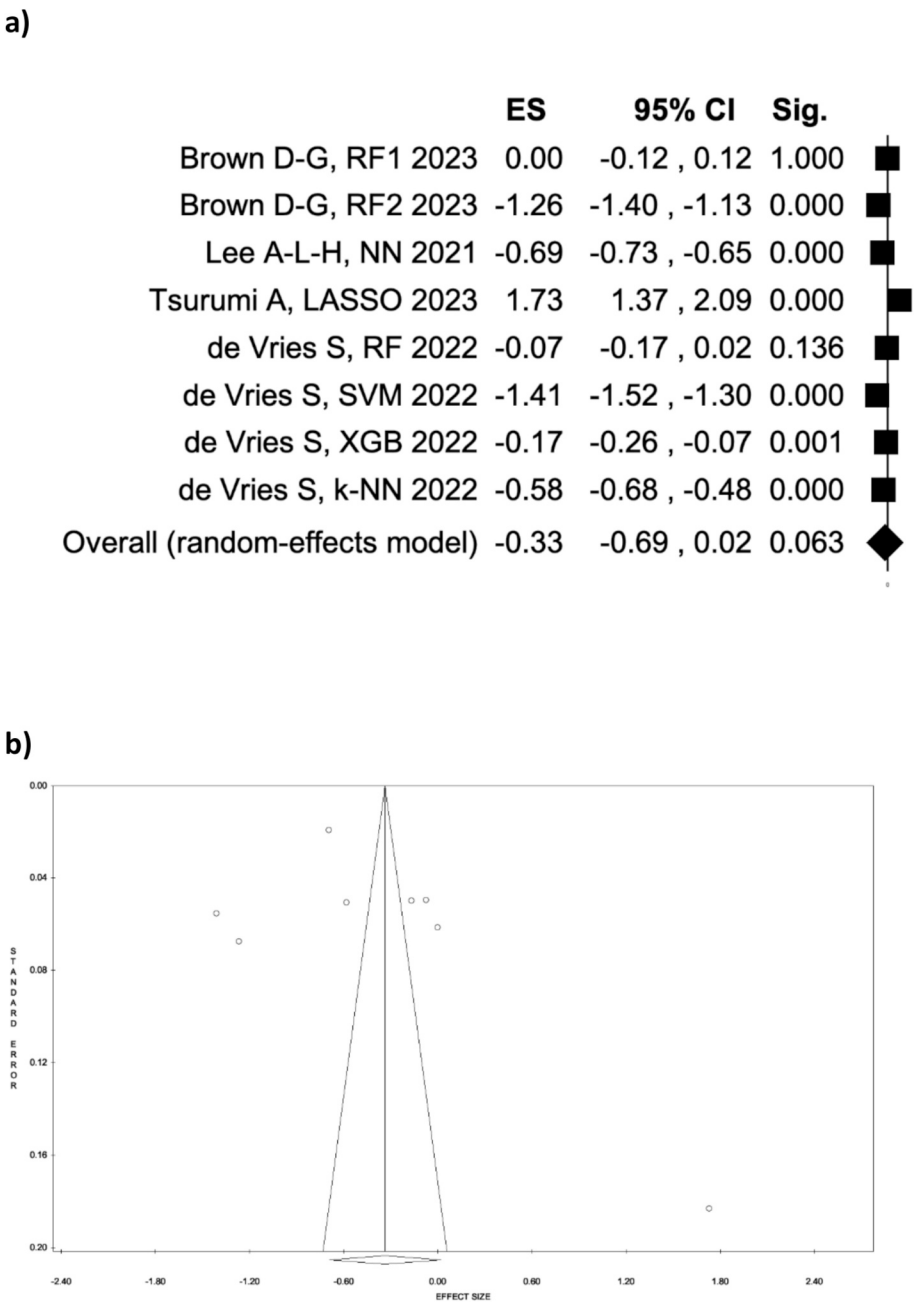


Fig. 8. a) forest plot and b) funnel plot of the random effect model assessing the ppv, with cohen's d as theeffect size (ES) measure.

overhead required for ML. Immediate replacement may create disruptions and inefficiencies, especially in healthcare systems unprepared for such a shift.

Lastly, the clinical validation of ML models across diverse settings is limited. Many models are developed and tested in controlled environments, which may not reflect the complexities and variations of real-world healthcare settings [69]. Until these tools are thoroughly validated externally and shown to consistently outperform traditional methods in diverse scenarios, their widespread adoption should be approached cautiously.

Prevention effort should also leverage ML capabilities to predict and monitor AMR trends in real time. Integrating ML tools into national and global surveillance programs might have the potential to enhance early detection of resistance patterns, informing targeted intervention strategies and optimizing resource allocation [70]. Furthermore, ML tools might also impact public health campaigns which can utilize insights derived from ML analyses to educate healthcare workers and the public about responsible antimicrobial use, amplifying the preventive impact of AMS initiatives [71–73].

In conclusion, while ML models represent a promising advancement in AMS, their immediate and complete replacement of traditional systems is premature. At the same time investments in developing standardized frameworks for the deployment and validation of ML algorithms across diverse healthcare settings are needed. In this perspective, guidelines for data sharing and model training to ensure equitable access and performance consistency are imperative. Additionally, policies should encourage collaborations between technology developers, healthcare providers, and policymakers to create user-friendly and interpretable ML solutions that align with clinical workflows. Establishing minimum standards for transparency and bias monitoring in ML models is critical to safeguard equitable access to high-quality AMS interventions.

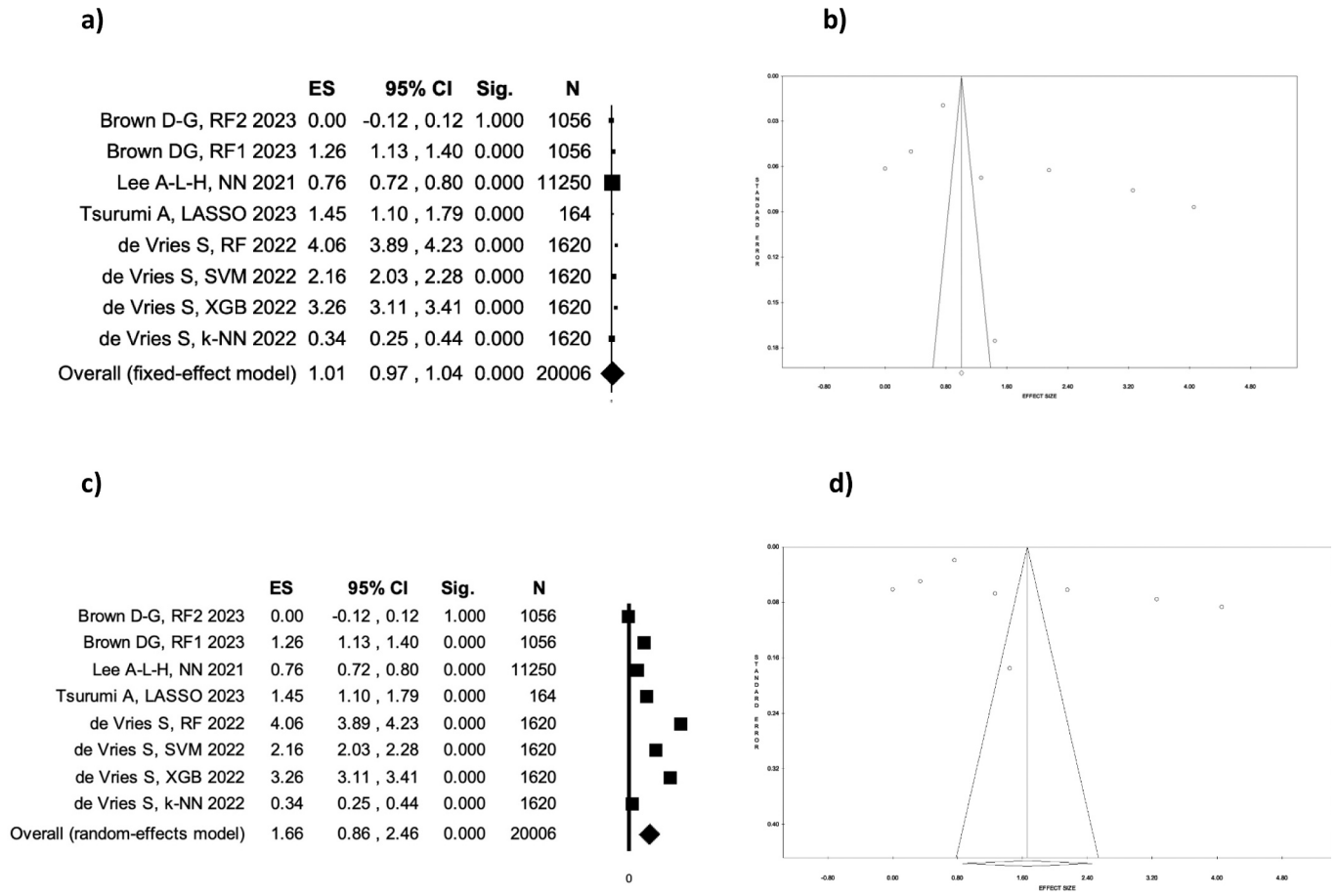


Fig. 9. a) forest plot and b) funnel plot of the fixed effect model and c) forest plot and d) funnel plot of the random effect model assessing the negative predictive value (npv), with cohen's d as theeffect size (ES) measure.

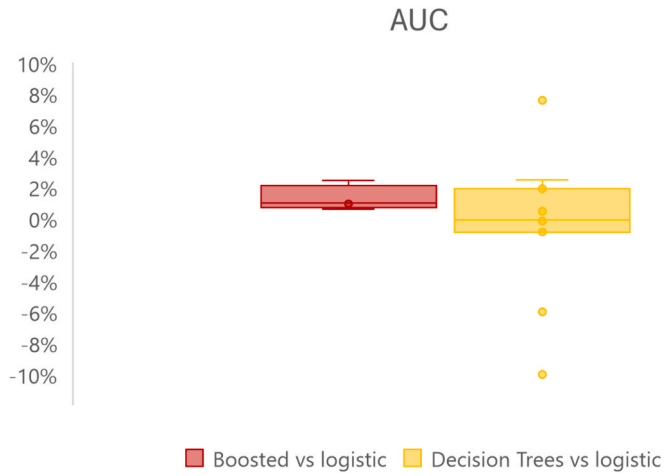


Fig. 10a. Sensitivity analysis for the AUC, comparing Boosted vs logistic (red) and decision trees vs logistic (yellow) when the random models are applied. The y-axis represents the difference in AUC between the alternative models (Boosted and DT) and the baseline model (LR). Positive values indicate an improvement in sensitivity relative to LR, while negative values suggest a decline. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

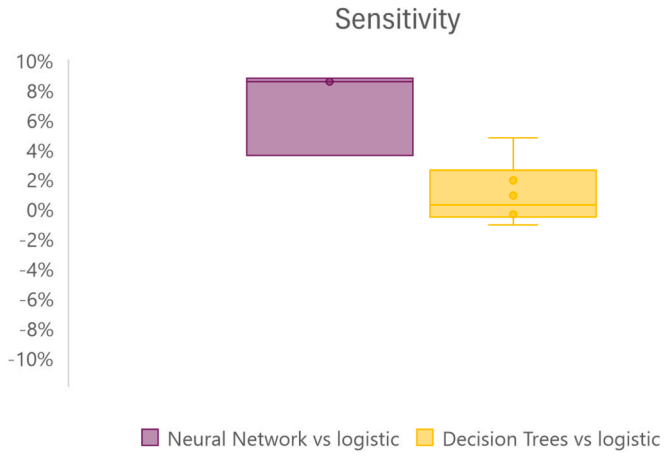


Fig. 10b. Sensitivity analysis for sensitivity, comparing NN vs logistic (purple) and decision trees vs logistic (yellow) when the random models are applied. The y-axis represents the difference in sensitivity between the alternative models (NN and DT) and the baseline model (LR). Positive values indicate an improvement in sensitivity relative to LR, while negative values suggest a decline. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.5. Strengths and limitations

This systematic review and meta-analysis provide key strengths, enhancing understanding of ML and traditional models in AMS. First, the inclusion of numerous studies across diverse ML algorithms and infection types improves generalizability. Second, a rigorous meta-analytic approach with standardized ES calculations and publication

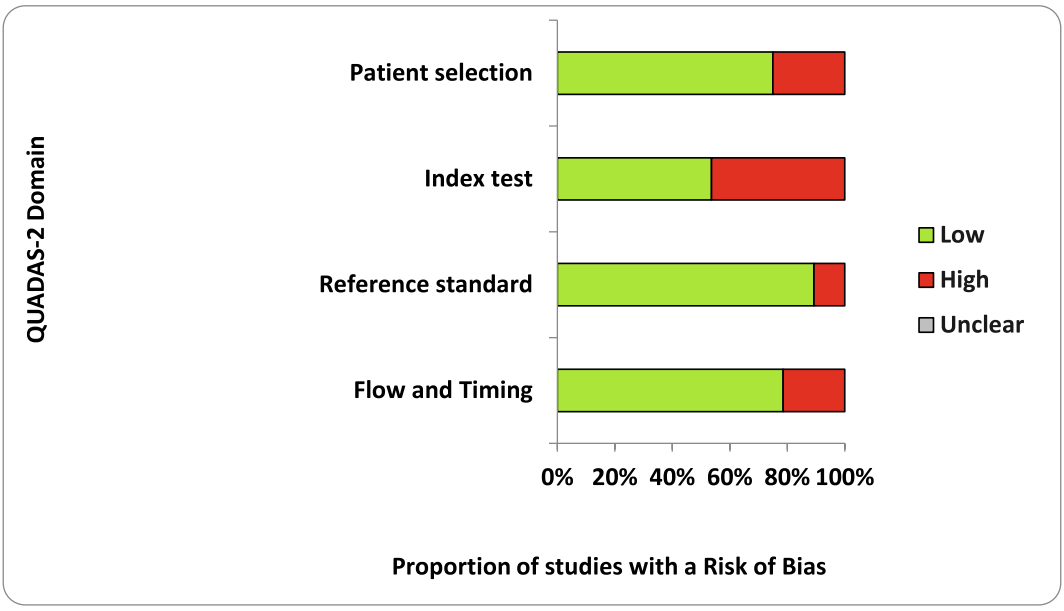


Fig. 11a. Proportion of studies with a risk of bias.

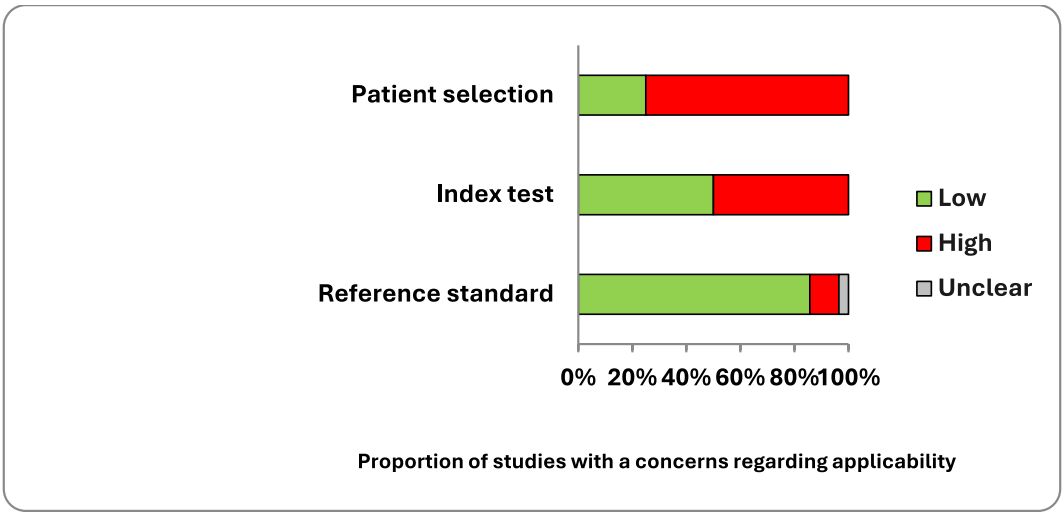


Fig. 11b. Proportion of studies with concerns regarding applicability.

bias assessments ensures robust evidence synthesis. Adherence to PRISMA guidelines and the PICOS framework ensures a methodologically sound and unbiased study selection process.

However, this review has several limitations. A high degree of heterogeneity was observed among the studies included, reflecting variations in study designs, populations, ML models, and measured outcomes. This heterogeneity extends across the diverse applications of ML in AMS, ranging from diagnostic microbiology and resistance detection to clinical decision support and epidemiological surveillance. While this broad scope highlights the versatility of ML, it also poses challenges in synthesizing findings and drawing direct comparisons due to variability in methodologies, data sources, and healthcare settings. To address this issue, we categorized the studies into five distinct groups based on the primary objective of ML application and conducted an additional sensitivity analysis. However, due to the inability to retrieve three articles and the lack of standard deviation data in some cases, the sensitivity analysis could only be performed for one out of the five groups. Despite these constraints, this approach aimed to enhance comparability and mitigate the impact of heterogeneity. Nonetheless, rather than being solely a limitation, this heterogeneity presents an opportunity to refine and tailor ML applications to specific AMS needs. A more standardized approach to model validation, reporting, and benchmarking would help mitigate these challenges and facilitate broader adoption in clinical practice. Future research should prioritize comparative effectiveness studies across different ML applications to identify the most clinically and operationally beneficial approaches in real-world settings.

Another limitation is the underrepresentation of certain patient populations, particularly outpatient and pediatric groups, which restricts the applicability of the results across all healthcare contexts. Additionally, many studies lacked detailed reporting on interpretability and calibration of ML models, which are critical factors for clinical implementation. Lastly, the exclusion of non-English studies could introduce language bias, potentially overlooking relevant evidence.

5. Conclusions

This systematic review and meta-analysis demonstrate that ML models hold significant promise for advancing AMS programs by

outperforming traditional methods in sensitivity and predictive accuracy. These advantages, especially in high-dimensional data environments, position ML as a transformative tool for addressing the global challenge of AMR. However, the findings also highlight critical gaps in external validation, interpretability, and equitable implementation that must be addressed to ensure the responsible integration of ML into clinical practice.

Current evidence underscores the need for a balanced approach that leverages ML's strengths while mitigating its risks. A cautious, phased adoption of ML in AMS, supported by investments in standardized frameworks for model development, validation, and deployment, is essential. Moreover, prioritizing explainable AI frameworks, enhancing data-sharing protocols, and focusing on underserved populations will be crucial to maximize the impact of ML on public health.

In conclusion, while ML has the potential to revolutionize AMS, its successful integration requires careful consideration of ethical, clinical, and infrastructural challenges. Future effort should focus on ensuring that ML models for AMS are not only accurate but also transparent, well-calibrated, externally validated, and equitably deployed, to fully realize their potential in addressing antimicrobial resistance on a global scale. Only through such a measured approach can ML fulfil its promise as a cornerstone of modern AMS strategies.

Ethical Statement

This study is a systematic review and meta-analysis based on previously published data. No new patient data were collected or analyzed, and no direct human or animal involvement occurred. Therefore, ethical approval was not required for this study. All included studies were conducted in accordance with ethical standards.

CRediT authorship contribution statement

Antonio Pinto: Conceptualization, Data curation, Methodology, Formal analysis, Investigation, Visualization, Writing – original draft. **Flavia Pennisi:** Conceptualization, Data curation, Methodology, Formal analysis, Investigation, Writing – original draft. **Giovanni Emanuele Ricciardi:** Methodology, Investigation, Writing – original draft. **Carlo Signorelli:** Conceptualization, Writing – review & editing, Supervision. **Vincenza Gianfredi:** Conceptualization, Project administration, Writing – original draft, Writing – review & editing, Supervision.

Funding

No funds, grants, or other support was received.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.idnow.2025.105090.

Data availability

The datasets generated and/or analyzed during the current study are available from the corresponding author upon reasonable request.

References

[1] AMR Review [Internet]. 2016. Tackling drug-resistant infections globally: final report and recommendations [Cited November 9, 2024]. Available from: https://amr-review.org/sites/default/files/160518_Final%20paper_with%20cover.pdf.
[2] Naghavi M, Vollset SE, Ikuta KS, Swetschinski LR, Gray AP, Wool EE, et al. Global burden of bacterial antimicrobial resistance 1990–2021: a systematic analysis with forecasts to 2050. Lancet 2024;404:1199–226. https://doi.org/10.1016/S0140-6736(24)01867-1.
[3] McGowan JE, Finland M. Usage of Antibiotics in a General Hospital: Effect of Requiring Justification. J Infect Dis 1974;130:165–8. https://doi.org/10.1093/infdis/130.2.165.
[4] Rice LB. Antimicrobial Stewardship and Antimicrobial Resistance. Med Clin N Am 2018;102:805–18. https://doi.org/10.1016/j.mcna.2018.04.004.
[5] Aiesh BM, Nazzal MA, Abdelhaq AI, Abutaha SA, Zyoud SH, Sabateen A. Impact of an antibiotic stewardship program on antibiotic utilization, bacterial susceptibilities, and cost of antibiotics. Sci Rep 2023;13:5040. https://doi.org/10.1038/s41598-023-32329-6.
[6] Weis CV, Jutzeler CR, Borgwardt K. Machine learning for microbial identification and antimicrobial susceptibility testing on MALDI-TOF mass spectra: a systematic review. Clin Microbiol Infect 2020;26:1310–7. https://doi.org/10.1016/j.cmi.2020.03.014.
[7] Pennisi F, Pinto A, Ricciardi GE, Signorelli C, Gianfredi V. The Role of Artificial Intelligence and Machine Learning Models in Antimicrobial Stewardship in Public Health: A Narrative Review. Antibiotics 2025;14:134. https://doi.org/10.3390/antibiotics14020134.
[8] Arterys [Internet]. n.d. Medical Imaging Cloud AI for Radiology [cited November 3, 2024]. Available from: https://www.arterys.com.
[9] Medtronic [Internet]. n.d. Medtronic home page [cited November 9, 2024]. Available from: https://www.medtronicdiabetes.com/.
[10] Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure F-X, et al. Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect 2020;26:584–95. https://doi.org/10.1016/j.cmi.2019.09.009.
[11] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. https://doi.org/10.1136/bmj.n71.
[12] Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak 2007;7:16. https://doi.org/10.1186/1472-6947-7-16.
[13] Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev 2016;5:210. https://doi.org/10.1186/s13643-016-0384-4. URL: https://rayyan.qcri.org.
[14] Nucci D, Santangelo OE, Provenzano S, Fatigoni C, Nardi M, Ferrara P, et al. Dietary Fiber Intake and Risk of Pancreatic Cancer: Systematic Review and Meta-Analysis of Observational Studies. Int J Environ Res Public Health 2021;18:11556. https://doi.org/10.3390/ijerph182111556.
[15] Sounderajah V, Ashrafian H, Rose S, Shah NH, Ghassemi M, Golub R, et al. A quality assessment tool for artificial intelligence-centered diagnostic test accuracy studies: QUADAS-AI. Nat Med 2021;27:1663–5. https://doi.org/10.1038/s41591-021-01517-0.
[16] Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front Psychol 2013;4:863. https://doi.org/10.3389/fpsyg.2013.00863.
[17] Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd edition. New York (United States): Routledge; 2013. 567 p. DOI: 10.4324/9780203771587.
[18] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557–60. https://doi.org/10.1136/bmj.327.7414.557.
[19] Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997;315:629–34. https://doi.org/10.1136/bmj.315.7109.629.
[20] Modongo C, Pasipanodya JG, Magazi BT, Srivastava S, Zetola NM, Williams SM, et al. Artificial Intelligence and Amikacin Exposures Predictive of Outcomes in Multidrug-Resistant Tuberculosis Patients. Antimicrob Agents Chemother 2016;60:5928–32. https://doi.org/10.1128/AAC.00962-16.
[21] Shah HN, Rajakaruna L, Ball G, Misra R, Al-Shahib A, Fang M, et al. Tracing the transition of methicillin resistance in sub-populations of Staphylococcus aureus, using SELDI-TOF Mass Spectrometry and Artificial Neural Network Analysis. Syst Appl Microbiol 2011;34:81–6. https://doi.org/10.1016/j.syapm.2010.11.002.
[22] Weis C, Cuénod A, Rieck B, Dubuis O, Graf S, Lang C, et al. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. Nat Med 2022;28:164–74. https://doi.org/10.1038/s41591-021-01619-9.
[23] Rawson TM, Hernandez B, Moore LSP, Herrero P, Charani E, Ming D, et al. A Real-world Evaluation of a Case-based Reasoning Algorithm to Support Antimicrobial Prescribing Decisions in Acute Care. Clin Infect Dis 2021;72:2103–11. https://doi.org/10.1093/cid/ciaa383.
[24] Mintz I, Chowers M, Obolski U. Prediction of ciprofloxacin resistance in hospitalized patients using machine learning. Commun Med 2023;3:43. https://doi.org/10.1038/s43856-023-00275-z.
[25] Kouchaki S, Yang Y, Walker TM, Sarah Walker A, Wilson DJ, Peto TEA, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics 2019;35:2276–82. https://doi.org/10.1093/bioinformatics/bty949.

[26] Gerada A, Harper N, Howard A, Reza N, Hope W. Determination of minimum inhibitory concentrations using machine-learning-assisted agar dilution. Microbiol Spectr 2024;12(5):e0420923. https://doi.org/10.1128/spectrum.04209-23.
[27] Kanjilal S, Sagers L, Li K, Byambajargal A, Sontag D. 527. Antimicrobial stewardship for empirical treatment of bloodstream infection using machine learning clinical decision support. Open Forum Infect Dis 2022;9. https://doi.org/10.1093/ofid/ofac492.582. ofac492 .582.
[28] Jones BE, Taber P, Ying J, Butler JM, Nevers M, Jones MM, et al. 1310. Provider and Facility Variation in Empiric Broad-Spectrum Antibiotic Use for Hospitalization Pneumonia: A Mixed Methods Study of Veterans Affairs Facilities. Open Forum Infect Dis 2021;8:S743–4. DOI: 10.1093/ofid/ofab466.1502.
[29] Huggins J, Hamilton KW, Barnett I. 768. Use of a Machine-Learning-based Prediction Model to Guide Antibiotic De-escalation in the Treatment of Urinary Tract Infections. Open Forum. Infect Dis 2019;6:S342–. https://doi.org/10.1093/ofid/ofz360.836.
[30] Ghemrawi N, Safi K, El Falou J. Assessment of Health Care Compliance in Managing Pseudomonas Aeruginosa in Urinary Tract Infection Using Machine Learning Techniques. 2023 Seventh International Conference on Advances in Biomedical Engineering (ICABME), IEEE; 2023:178–83. DOI: 10.1109/ICABME59496.2023.10293106.
[31] Black CA, Aguilar S, Bandy S, Gawrys G, Dallas S, So W, et al. 118. Machine Learning Approaches to Predicting Treatment Outcomes for Carbapenem-Resistant Enterobacterales in a Region with High Prevalence of Non-Carbapenemase Producers. Open Forum. Infect Dis 2021;8:S71–. https://doi.org/10.1093/ofid/ofab466.118.
[32] Dang J, Shu J, Wang R, Yu H, Chen Z, Yan W, et al. The glycopatterns of *Pseudomonas aeruginosa* as a potential biomarker for its carbapenem resistance. Microbiol Spectr 2023;11(6):e0200123. https://doi.org/10.1128/spectrum.02001-23.
[33] Rezaei-hachesu P, Samad-Soltani T, Yaghoubi S, GhaziSaeedi M, Mirnia K, Masoumi-Asl H, et al. The design and evaluation of an antimicrobial resistance surveillance system for neonatal intensive care units in Iran. Int J Med Inform 2018;115:24–34. https://doi.org/10.1016/j.ijmedinf.2018.04.007.
[34] Bhavani SV, Lonjers Z, Carey KA, Afshar M, Gilbert ER, Shah NS, et al. The Development and Validation of a Machine Learning Model to Predict Bacteremia and Fungemia in Hospitalized Patients Using Electronic Health Record Data. Crit Care Med 2020;48:e1020–8. https://doi.org/10.1097/CCM.0000000000004556.
[35] Brown DG, Worby CJ, Pender MA, Brintz BJ, Ryan ET, Sridhar S, et al. Development of a prediction model for the acquisition of extended spectrum beta-lactam-resistant organisms in U.S. international travellers. J Travel Med 2023;30(6). https://doi.org/10.1093/jtm/taad028. taad028.
[36] Çağlayan Ç, Barnes SL, Pineles LL, Harris AD, Klein EY. A Data-Driven Framework for Identifying Intensive Care Unit Admissions Colonized With Multidrug-Resistant Organisms. Front Public Health 2022;10:853757. https://doi.org/10.3389/fpubh.2022.853757.
[37] Goodman KE, Heil EL, Claeys KC, Banoub M, Bork JT. Real-world antimicrobial stewardship experience in a large academic medical center: using statistical and machine learning approaches to identify intervention "hotspots" in an antibiotic audit and feedback program. Open Forum Infect Dis 2022;9(7). https://doi.org/10.1093/ofid/ofac289.
[38] Nigo M, Rasmy L, Mao B, Kannadath BS, Xie Z, Zhi D. Deep learning model for personalized prediction of positive MRSA culture using time-series electronic health records. Nat Commun 2024;15:2036. https://doi.org/10.1038/s41467-024-46211-0.
[39] Shang JS, Lin YE, Goetz AM. Diagnosis of MRSA with neural networks and logistic regression approach. Health Care Manag Sci 2000;3:287–97. https://doi.org/10.1023/A:1019018129822.
[40] Tsurumi A, Flaherty PJ, Que Y-A, Ryan CM, Banerjee A, Chakraborty A, et al. A Preventive tool for predicting bloodstream infections in children with burns. Shock 2023;59:393–9. https://doi.org/10.1097/SHK.0000000000002075.
[41] Garcia-Vidal C, Puerta-Alcalde P, Cardozo C, Orellana MA, Besanson G, Lagunas J, et al. Machine learning to assess the risk of multidrug-resistant gram-negative bacilli infections in febrile neutropenic hematological patients. Infect Dis Ther 2021;10:971–83. https://doi.org/10.1007/s40121-021-00438-2.
[42] Jiménez F, Palma J, Sánchez G, Marín D, Francisco Palacios MD, Lucía López MD. Feature selection based multivariate time series forecasting: An application to antibiotic resistance outbreaks prediction. Artif Intell Med 2020;104:101818. https://doi.org/10.1016/j.artmed.2020.101818.
[43] Martínez-Agüero S, Mora-Jiménez I, Lérida-García J, Álvarez-Rodríguez J, Soguero-Ruiz C. Machine Learning Techniques to Identify Antimicrobial Resistance in the Intensive Care Unit. Entropy 2019;21:603. https://doi.org/10.3390/e21060603.
[44] Huang T-S, Lee S-S-J, Lee C-C, Chang F-C. Detection of carbapenem-resistant Klebsiella pneumoniae on the basis of matrix-assisted laser desorption ionization time-of-flight mass spectrometry by using supervised machine learning approach. PLoS One 2020;15:e0228459. https://doi.org/10.1371/journal.pone.0228459.
[45] Tsai W-C, Liu C-F, Ma Y-S, Chen C-J, Lin H-J, Hsu C-C, et al. Real-time artificial intelligence system for bacteremia prediction in adult febrile emergency department patients. Int J Med Inform 2023;178:105176. https://doi.org/10.1016/j.ijmedinf.2023.105176.
[46] Wang C, Wang Z, Wang H-Y, Chung C-R, Horng J-T, Lu J-J, et al. Large-Scale Samples Based Rapid Detection of Ciprofloxacin Resistance in Klebsiella pneumoniae Using Machine Learning Methods. Front Microbiol 2022;13:827451. https://doi.org/10.3389/fmicb.2022.827451.
[47] Liang Q, Zhao Q, Xu X, Zhou Y, Huang M. Early prediction of carbapenem-resistant Gram-negative bacterial carriage in intensive care units using machine learning. J Glob Antimicrob Resist 2022;29:225–31. https://doi.org/10.1016/j.jgar.2022.03.019.
[48] Liang Q, Ding S, Chen J, Chen X, Xu Y, Xu Z, et al. Prediction of carbapenem-resistant gram-negative bacterial bloodstream infection in intensive care unit based on machine learning. BMC Med Inform Decis Mak 2024;24:123. https://doi.org/10.1186/s12911-024-02504-4.
[49] Beaudoin M, Kabanza F, Nault V, Valiquette L. Evaluation of a machine learning capability for a clinical decision support system to enhance antimicrobial stewardship programs. Artif Intell Med 2016;68:29–36. https://doi.org/10.1016/j.artmed.2016.02.001.
[50] Bystritsky RJ, Beltran A, Young AT, Wong A, Hu X, Doernberg SB. Machine learning for the prediction of antimicrobial stewardship intervention in hospitalized patients receiving broad-spectrum agents. Infect Control Hosp Epidemiol 2020;41:1022–7. https://doi.org/10.1017/ice.2020.213.
[51] de Vries S, ten Doesschate T, Totté JEE, Heutz JW, Loeffen YGT, Oosterheert JJ, et al. A semi-supervised decision support system to facilitate antibiotic stewardship for urinary tract infections. Comput Biol Med 2022;146:105621. https://doi.org/10.1016/j.compbiomed.2022.105621.
[52] Eickelberg G, Sanchez-Pinto LN, Luo Y. Predictive modeling of bacterial infections and antibiotic therapy needs in critically ill adults. J Biomed Inform 2020;109:103540. https://doi.org/10.1016/j.jbi.2020.103540.
[53] Feretzakis G, Loupelis E, Sakagianni A, Kalles D, Martsoukou M, Lada M, et al. Using Machine Learning Techniques to Aid Empirical Antibiotic Therapy Decisions in the Intensive Care Unit of a General Hospital in Greece. Antibiotics 2020;9:50. https://doi.org/10.3390/antibiotics9020050.
[54] İlhanlı N, Park SY, Kim J, Ryu JA, Yardımcı A, Yoon D. Prediction of Antibiotic Resistance in Patients With a Urinary Tract Infection: Algorithm Development and Validation. JMIR Med Inform 2024;12:e51326. https://doi.org/10.2196/51326.
[55] Oonsivilai M, Mo Y, Luangasanatip N, Lubell Y, Miliya T, Tan P, et al. Using machine learning to guide targeted and locally-tailored empiric antibiotic prescribing in a children's hospital in Cambodia. Wellcome Open Res 2018;3:131. https://doi.org/10.12688/wellcomeopenres.14847.1.
[56] Wong JG, Aung A-H, Lian W, Lye DC, Ooi C-K, Chow A. Risk prediction models to guide antibiotic prescribing: a study on adult patients with uncomplicated upper respiratory tract infections in an emergency department. Antimicrob Resist Infect Control 2020;9:171. https://doi.org/10.1186/s13756-020-00825-3.
[57] Ananda-Rajah MR, Bergmeir C, Petitjean F, Slavin MA, Thursky KA, Webb GI. Toward electronic surveillance of invasive mold diseases in hematology-oncology patients: an expert system combining natural language processing of chest computed tomography reports, microbiology, and antifungal drug data. JCO Clin Cancer Inform 2017;1:1–10. https://doi.org/10.1200/CCI.17.00011.
[58] Herman B, Sirichokchatchawan W, Pongpanich S, Nantasenamat C. Development and performance of CUHAS-ROBUST application for pulmonary rifampicin-resistance tuberculosis screening in Indonesia. PLoS One 2021;16:e0249243. https://doi.org/10.1371/journal.pone.0249243.
[59] Lee ALH, To CCK, Lee ALS, Chan RCK, Wong JSH, Wong CW, et al. Deep learning model for prediction of extended-spectrum beta-lactamase (ESBL) production in community-onset Enterobacteriaceae bacteraemia from a high ESBL prevalence multi-centre cohort. Eur J Clin Microbiol Infect Dis 2021;40:1049–61. https://doi.org/10.1007/s10096-020-04120-2.
[60] Tacconelli E, Górska A, De Angelis G, Lammens C, Restuccia G, Schrenzel J, et al. Estimating the association between antibiotic exposure and colonization with extended-spectrum β-lactamase-producing Gram-negative bacteria using machine learning methods: a multicentre, prospective cohort study. Clin Microbiol Infect 2020;26:87–94. https://doi.org/10.1016/j.cmi.2019.05.013.
[61] Pennisi F, Pinto A, Ricciardi GE, Signorelli C, Gianfredi V. Artificial intelligence in antimicrobial stewardship: a systematic review and meta-analysis of predictive performance and diagnostic accuracy. Eur J Clin Microbiol Infect Dis 2025;44(3):463–513. https://doi.org/10.1007/s10096-024-05027-y.
[62] Benti NE, Chaka MD, Semie AG. Forecasting Renewable Energy Generation with Machine Learning and Deep Learning: Current Advances and Future Prospects. Sustainability 2023;15:7087. https://doi.org/10.3390/su15097087.
[63] Yoon CH, Torrance R, Scheinerman N. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? J Med Ethics 2022;48:581–5. https://doi.org/10.1136/medethics-2020-107102.
[64] Myllyaho L, Raatikainen M, Männistö T, Mikkonen T, Nurminen JK. Systematic literature review of validation methods for AI systems. J Syst Softw 2021;181:111050. https://doi.org/10.1016/j.jss.2021.111050.
[65] Gala D, Behl H, Shah M, Makaryus AN. The Role of Artificial Intelligence in Improving Patient Outcomes and Future of Healthcare Delivery in Cardiology: A Narrative Review of the Literature. Healthcare 2024;12:481. https://doi.org/10.3390/healthcare12040481.
[66] Signorelli C, Pennisi F, Lunetti C, Blandi L, Pellissero G. Fondazione Sanità Futura WG. Quality of hospital care and clinical outcomes: a comparison between the Lombardy Region and the Italian national data. Ann Ig 2024;36:234–49. https://doi.org/10.7416/ai.2024.2597.
[67] Li G, Hu Y, Chen H, Li H, Hu M, Guo Y, et al. A sensor fault detection and diagnosis strategy for screw chiller system using support vector data description-based D-statistic and DV-contribution plots. Energy Build 2016;133:230–45. https://doi.org/10.1016/j.enbuild.2016.09.037.
[68] Hofer IS, Burns M, Kendale S, Wanderer JP. Realistically Integrating Machine Learning Into Clinical Practice: A Road Map of Opportunities, Challenges, and a Potential Future. Anesth Analg 2020;130:1115–8. https://doi.org/10.1213/ANE.0000000000004575.
[69] Karalis VD. The Integration of Artificial Intelligence into Clinical Practice. Appl Biosci 2024;3:14–44. https://doi.org/10.3390/applbiosci3010002.

[70] Signorelli C, De Ponti E, Mastrangelo M, Pennisi F, Cereda D, Corti F, et al. The contribution of the private healthcare sector during the COVID-19 pandemic: the experience of the Lombardy Region in Northern Italy. Ann Ig 2024;36:250–5. https://doi.org/10.7416/ai.2024.2609.
[71] Pennisi F, Minerva M, Dalla Valle Z, Odone A, Signorelli C. Genesis and prospects of the shortage of specialist physicians in Italy and indicators of the 39 schools of hygiene and preventive medicine. Acta Biomed 2023;94(S3):e2023159. https://doi.org/10.23750/abm.v94iS3.14512.
[72] Pennisi F, Genovese C, Gianfredi V. Lessons from the COVID-19 Pandemic: Promoting Vaccination and Public Health Resilience, a Narrative Review. Vaccines (Basel) 2024;12(8):891. https://doi.org/10.3390/vaccines12080891.
[73] Signorelli C, Pennisi F, D'Amelio AC, Conversano M, Cinquetti S, Blandi L, et al. Vaccinating in Different Settings: Best Practices from Italian Regions. Vaccines (Basel) 2024;13(1):16. https://doi.org/10.3390/vaccines13010016.