World Journal of Urology (2024) 42:464
https://doi.org/10.1007/s00345-024-05145-4

RESEARCH

Artificial intelligence and machine learning applications in urinary tract infections identification and prediction: a systematic review and meta-analysis

Li Shen[1] · Jialu An[2] · Nanding Wang[3] · Jin Wu[4] · Jia Yao[5,6] · Yumei Gao[1]

Received: 16 February 2024 / Accepted: 23 June 2024 / Published online: 1 August 2024

Abstract

Background Urinary tract infections (UTIs) have been one of the most common bacterial infections in clinical practice worldwide. Artificial intelligence (AI) and machine learning (ML) based algorithms have been increasingly applied in UTI case identification and prediction. However, the overall performance of AI/ML algorithms in identifying and predicting UTI has not been evaluated. The purpose of this paper is to quantitatively evaluate the application value of AI/ML in identifying and predicting UTI cases.

Methods MEDLINE, EMBASE, Web of Science, and PubMed databases were systematically searched for articles published up to December 31, 2023. Quality Assessment of Diagnostic Accuracy Studies tool (QUADAS-2) and Prediction Model Risk of Bias Assessment Tool (PROBAST) were used to assess the risk of bias. Study characteristics and detailed algorithm information were extracted. Pooled sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) were synthesized using a bivariate mix-effects model. Meta-regression and subgroup analysis were conducted to test the source of heterogeneity.

Results In total, 11 studies with 14 AI/ML models were included in the final meta-analysis. The overall pooled AUC was 0.89 (95%CI 0.86–0.92). Additionally, the pooled Sen, Spe, PLR, NLR, and DOR were 0.78 (95%CI 0.71–0.84), 0.89 (95%CI 0.83–0.93), 6.99 (95%CI 4.38–11.14), 0.25 (95%CI 0.18–0.34) and 28.07 (95%CI 14.27–55.20), respectively. The results of meta-regression suggested that reference standard definitions might be the source of heterogeneity.

Conclusion AI/ML algorithms appear to be promising to help clinicians detect and identify patients at high risk of UTIs. However, further studies are demanded to evaluate the application value of AI/ML more thoroughly.

Keywords Artificial intelligence · Machine learning · Urinary tract infections · Meta-analysis · Prediction model · Diagnostic accuracy

✉ Yumei Gao
gaoyumei_zy@163.com

Li Shen
shenli_cmu@163.com

Jialu An
an_71@163.com

Nanding Wang
93404582@qq.com

Jin Wu
493116097@qq.com

Jia Yao
282652404@qq.com

[1] Department of Infection Control, Xi'an Hospital of Traditional Chinese Medicine, No.69 Feng Cheng 8th Road, Weiyang District, Xi'an 710021, China

[2] Department of Information Consultation, Library of Xi'an Jiaotong University, No.76 Yan Ta West Road, Yanta District, Xi'an 710061, China

[3] Department of Cardiology, Xi'an Hospital of Traditional Chinese Medicine, No.69 Feng Cheng 8th Road, Weiyang District, Xi'an 710021, China

[4] Department of Clinical Laboratory, Xi'an Hospital of Traditional Chinese Medicine, No.69 Feng Cheng 8th Road, Weiyang District, Xi'an 710021, China

[5] Experimental Center, Xi'an Hospital of Traditional Chinese Medicine, No.69 Feng Cheng 8th Road, Weiyang District, Xi'an 710021, China

[6] Xi'an Academy of Traditional Chinese Medicine, No.69 Feng Cheng 8th Road, Weiyang District, Xi'an 710021, China

Abbreviations

AI	Artificial intelligence
ML	Machine learning
UTI	Urinary tract infections
CAUTI	Catheter-associated urinary tract infections
USD	United States dollar
HAI	Healthcare-associated infections
EMR	Electronic medical records
PROSPERO	International Prospective Register of Systematic Reviews
CRD	Centre of Reviews and Dissemination
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analysis
TP	True positive
FP	False positive
FN	False negative
TN	True negative
AUC	Area under the receiver operating characteristic curve
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
PROBAST	Prediction Model Risk of Bias Assessment Tool
PLR	Positive likelihood ratio
NLR	Negative likelihood ratio
DOR	Diagnostic odds ratio
CI	Confidence interval
SROC	Summary of receiver operating characteristics
LR	Logistic regression
RF	Random forest
ANN	Artificial neural network
XGBoost	Extreme gradient boosting
SVM	Support vector machine
SMOTETomek	Synthetic Minority Oversampling Technique and Tomek Links Undersampling
LASSO	Least absolute shrinkage and selection operator
ICD	International Classification of Diseases
ACS-NSQIP	American College of Surgeons National Surgical Quality Improvement Program
EAU	European Association of Urology
POETIC	Point of care testing for urinary infection in primary care
HAIBA	Hospital-Acquired Infections Database
HAIR	Hospital Acquired Infection Registry
CDC	Centers for Disease Control and Prevention

Background

Urinary tract infections (UTIs) are among the most common bacterial infections in clinical practice, both in the community and in hospitals. Annually, 150 million people acquire UTIs worldwide [1]. The incidence of UTIs is generally higher in women than in men. It is estimated that more than 50% of women will experience a UTI at least once during their lifetime, of whom 20–30% will have recurrent UTIs in the future [2]. In the United States, an estimated 1 million cases of hospital-acquired UTI occur per year, 80% of which are attributable to indwelling urinary catheters [3]. Catheter-associated UTIs (CAUTIs) have a significant impact on increased mortality, prolonged hospital length of stay, serious antimicrobial resistance and extra health care costs. Each symptomatic hospital-acquired UTI episode costs an additional $676, and urinary catheter-related bacteremia increases expenses by $2836 [4]. In China, UTIs have been one of the most frequent healthcare-associated infections (HAIs) in general hospitals, accounting for 11.65% of HAIs [5]. In 2019, there were 19.62 million incident UTI cases among females and 2.76 million among males. The 50 ~ 69 year age group had the highest incidence of UTIs, at 2.27%, representing a significant health care burden [6].

UTIs include pathogenic infections at anatomical sites from the urethra to the kidney. Currently, the diagnosis of UTIs is usually based on a combination of clinical presentations and positive urine analysis or culture. The gold standard for diagnosing UTIs is the detection and identification of the pathogen in relevant quantities through urine cultures [7]. However, this process typically takes 24 to 48 h for complete isolation and identification of the pathogen. Before the results of urine culture are available, empirical antibiotic treatment is often necessary to relieve symptoms, shorten their duration, and prevent disease progression. However, it has been discovered that approximately 40 to 42% of female patients visiting an emergency department diagnosed with UTIs by clinicians had unlikely or incorrect diagnoses [8]. Misdiagnosis and overdiagnosis can lead to excessively unnecessary prescriptions, contributing to the increasing development of antibiotic resistance. Therefore, accurate identification and prediction of UTIs before urine cultures available can be crucial in guiding prescription practices and preparing for future resistance threats.

In recent years, artificial intelligence (AI) has been suggested to be a revolutionary strategy that creates predictive tools providing diagnostic assistance for clinical decision-making. As a subset of AI, machine learning (ML) allows the construction of an algorithm that can learn, identify patterns, recognize relationships, and predict outcomes automatically through training and testing data, primarily from electronic medical records (EMR) datasets. AI/ML algorithms have

been increasingly applied in UTI case identification and prediction [9]. However, the effectiveness of these AI/ML algorithms has not been evaluated and summarized. To our knowledge, there is no evidence-based research that quantitatively evaluates the predictive performance of these AI/ML algorithms in UTI diagnosis and discrimination. In this study, we conducted a systematic review and meta-analysis to assess the diagnostic accuracy and reliability of AI/ML algorithms for UTI, aiming to provide insights for future directions.

Methods

Ethics approval for this study was not required. The systematic review protocol was previously registered on PROSPERO (International Prospective Register of Systematic Reviews) with the report number CRD42023431512. The Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement was followed.

Search strategy

A comprehensive systematic literature search was performed in MEDLINE, EMBASE, Web of Science, and PubMed databases for articles published up to 31st December, 2023. The first Boolean search was conducted to identify the focused AI/ML algorithms by using the term "or" to explode the following subject headings: "machine learning", "artificial intelligence", "deep learning", "neural networks", "computer-assisted", "data mining" and "algorithms". The second Boolean search was conducted by using the term "or" to explode the subject headings: "urinary tract infection", "bacteriuria". The above two Boolean searches were combined by the Boolean term "and". Language was not limited at the time of search. The detailed search strategy was presented in Supplementary Table S1. We also manually searched the reference lists of included studies to identify any relevant articles.

Eligibility criteria

The inclusion criteria were as follows: (1) studies evaluating the diagnostic accuracy of UTI; (2) studies that used AI/ML algorithms; (3) participants aged 18 years or older; (4) studies from which a 2×2 confusion matrix can be extracted or reconstructed. Excluded criteria were non-human studies, case reports, editorials, protocols, comment or letters, and reviews.

AI/ML algorithms

Logistic Regression (LR): LR models the probability of a binary outcome by fitting data to a logistic curve, offering interpretability and suitability for linearly separable data, though it struggles with complex non-linear patterns.

Support vector machine (SVM): SVM is a valuable tool for classification tasks with small sample sizes. SVM works by finding the optimal hyperplane that best separates different classes of data points in high-dimensional spaces.

Decision tree (DT): DT is like a flowchart where data is split into different categories based on a series of questions. It's easy to understand and can handle complex data. Yet, it's susceptible to overfitting, which might compromise its performance on new data.

Random forest (RF): RF builds multiple decision trees during training and aggregate outputs to make predictions, reducing overfitting and improving robustness, but at the expense of model interpretability and longer training times.

Artificial neural networks (ANN): ANN replicate the structure and functioning of the human brain, comprising interconnected nodes (neurons) organized in layers, where each neuron processes inputs and passes outputs to the next layer. They can process complex non-linear relationships and large datasets efficiently. But they require abundant data to avoid overfitting.

Boosting algorithms: Boosting algorithms work by sequentially training models, with each new model correcting the errors of its predecessor. The process continues until a predefined stopping criteria is met. There are several types of Boosting algorithms, such as AdaBoost, Gradient Boosting, and XGBoost, known for their high predictive accuracy.

Ensemble algorithms: Ensemble algorithms leverage the idea that aggregating predictions from multiple models often leads to better results than using a single model. Common ensemble methods include bagging, boosting, and stacking. They are popular due to superior predictive performance across different types of data.

Study selection

Two research authors (LS and JA) independently examined the titles and abstracts of the records retrieved based on the selection criteria. Full-text articles were subsequently screened after duplicates were removed. If a title or abstract could not be judged, the full text was read to decide whether the article should be included or not. Discrepancies in inclusion studies were discussed and resolved by consensus.

Data extraction

Data extraction was performed independently by two research authors (LS and JA) using a standardized form, which consisted of author, publication year, country of origin, study design and setting, participants, sample size, percentage of female patients, details of AI/ML algorithms, reference standard, data sources, input variables and outcomes. For eligible studies in the final meta-analysis, we also extracted the diagnostic accuracy results data including the number of true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN), sensitivity, specificity, and area under the receiver operating characteristic curve (AUC) predicted by the best-performing AI/ML model in the validation (internal or external) datasets. When studies did not report exact values for TP, FP, FN, TN, we manually calculated and reconstructed confusion matrix using the total sample size (N), prevalence (P), sensitivity (Sen), and specificity (Spe). And the formula were: TP = N×P×Sen, FN = N×P-TP, TN = N×(1-P)×Spe, FP = N×(1-P)-TN. Any disagreement during this process was resolved by consulting a third author (NW).

Risk of bias assessment

A Quality Assessment of Diagnostic Accuracy Studies tool [10] (QUADAS-2) was used to evaluate the quality of the included studies and potential bias by two independent authors (JA and JY). We further performed quality and risk of bias assessment using Prediction Model Risk of Bias Assessment Tool [11] (PROBAST). Disagreements were solved with consensus by a third researcher (NW). Deeks' funnel plot was applied to assess publication bias using an asymmetry test by Stata 15.0 software. If $P < 0.05$, publication bias may be present.

Statistical analysis

To evaluate the diagnostic performance of AI/ML models, the Sen, Spe, positive likelihood ratio (PLR), negative likelihood ratio (NLR) and diagnostic odds ratio (DOR) along with a 95% confidence interval (CI) were calculated separately for each AI/ML algorithm. A bivariate mix-effects model was applied in the calculation of the pooled results considering suspected high proportion of heterogeneity. Then forest plots showing the separate and summary results of sensitivity and specificity were generated. Summary of receiver operating characteristics (SROC) curves were adopted to assess overall diagnostic accuracy and the SROC's AUC was calculated. The heterogeneity between included studies was quantified by the I^2 statistics and Cochran Q statistic. I^2 exceeding 50% and $P < 0.05$ indicated potential heterogeneity. Sensitivity analysis was performed to evaluate the robustness of combined results by sequentially removing each individual study one at a time from the analysis. Subgroup analysis and meta-regression were conducted in order to explore any source of heterogeneity. Heterogeneity due to threshold effect was considered when the Spearman correlation coefficient value was great than 0.5 with $P < 0.05$. A 2-sided P value < 0.05 indicated statistical significance for all tests. Statistical analysis was performed using Stata 15.0, Meta-DiSc 1.4, and RevMan 5.3 software.

Results

Study selection

A total of 4595 records were retrieved across the electronic databases and via manual search. After eliminating duplicates and ineligible article type, we screened 3712 records by titles and abstracts. Among these, 71 articles of interest were assessed for eligibility through full-text reading. Twenty papers [12–31] met the predefined inclusion criteria for systematic review. We then conducted a meta-analysis on 11 articles [12–16, 19, 22, 25, 27, 29, 31] with available data. The specific study selection process and reasons for exclusion are shown in Fig. 1.

Study characteristics

All 20 included articles were retrospective observational studies that examined the applications of AI/ML for identifying and predicting UTIs. A total of 3,032,281 participants were involved in these studies, with 54.65% being female. The mean age of participants ranged from 41 to 64 years. The majority of the selected literature was published within the past six years, with more than half being published between 2022–2023 [12, 13, 15, 16, 20, 22, 24, 25, 28, 30]. There were 14 multi-center studies and 6 single-center studies. Twenty-one AI/ML algorithms were adopted in these studies. Among them, 75% employed cross-validation techniques, while 30% utilized external validation datasets for model evaluation. The proportions of studies using LR, RF, ANN, XGBoost, SVM and DT modeling were 60%, 60%, 35%, 35%, 25% and 25% respectively. The optimal prediction model for identifying UTI was constructed using ElasticNet algorithm, achieving an AUC of 0.94 [14]. The characteristics of the included studies are listed in Supplementary Table S2.

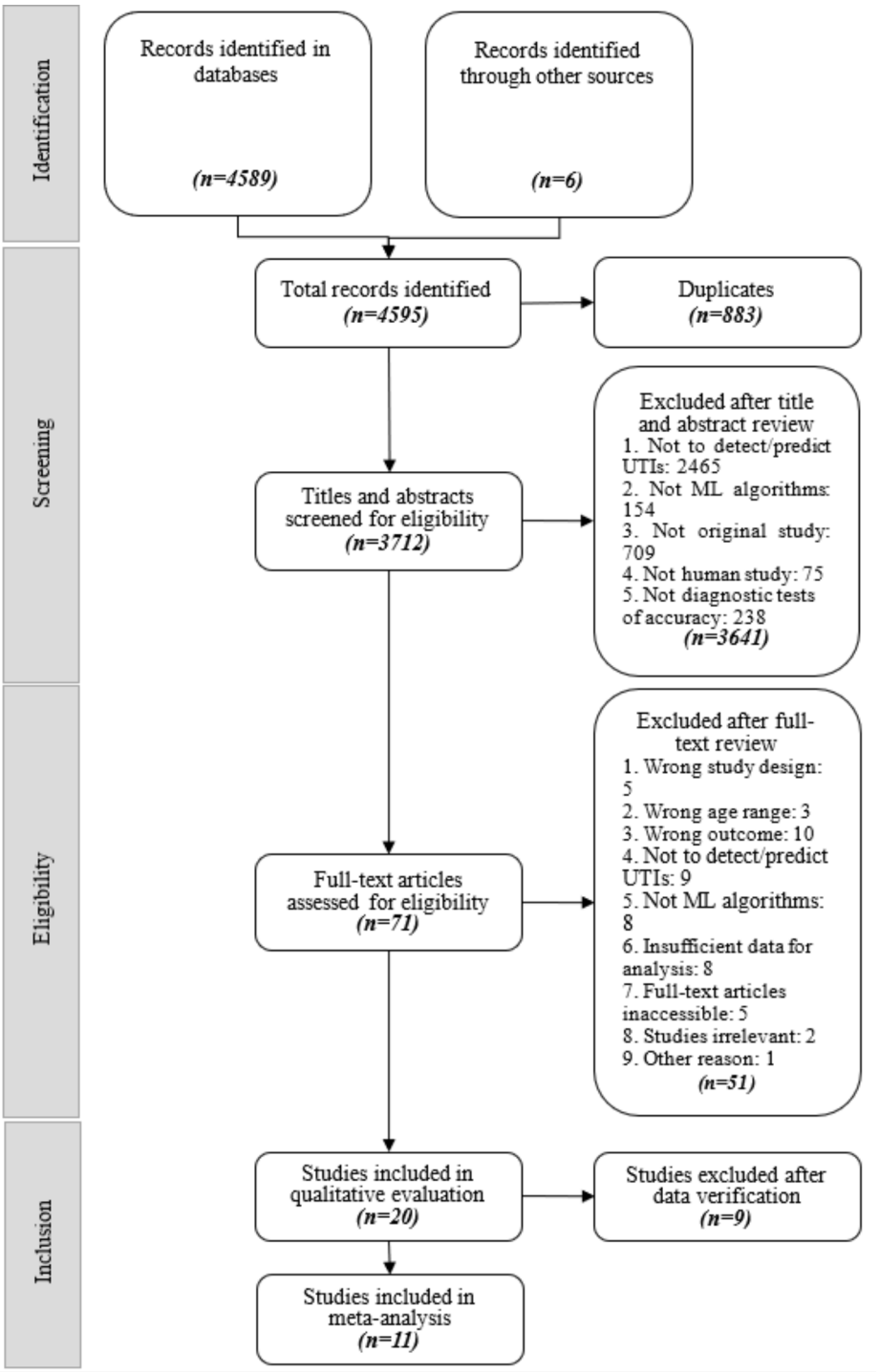

Fig. 1 PRISMA flow diagram of study selection process

Quality assessment and publication bias

Regarding the QUADAS-2 tool, the quality assessment results of included studies are summarized in Fig. 2. Bias in the included studies primarily stems from the domains of patient selection and index test. The main reasons for the risk of bias are displayed in Supplementary Table S3. Of all the included studies, only 3 articles were judged as "low risk" on all domains. In terms of overall concerns regarding applicability, 9 out of 20 presented "low concern". Furthermore, we evaluated the risk of bias using PROBAST. The assessment results showed that none of the included studies were judged as "low risk" on all domains. The main source of high risk of bias lied in the insufficient number of participants with the outcome (60.00%). Unclear risk of bias mainly originated from the unclear correlation between

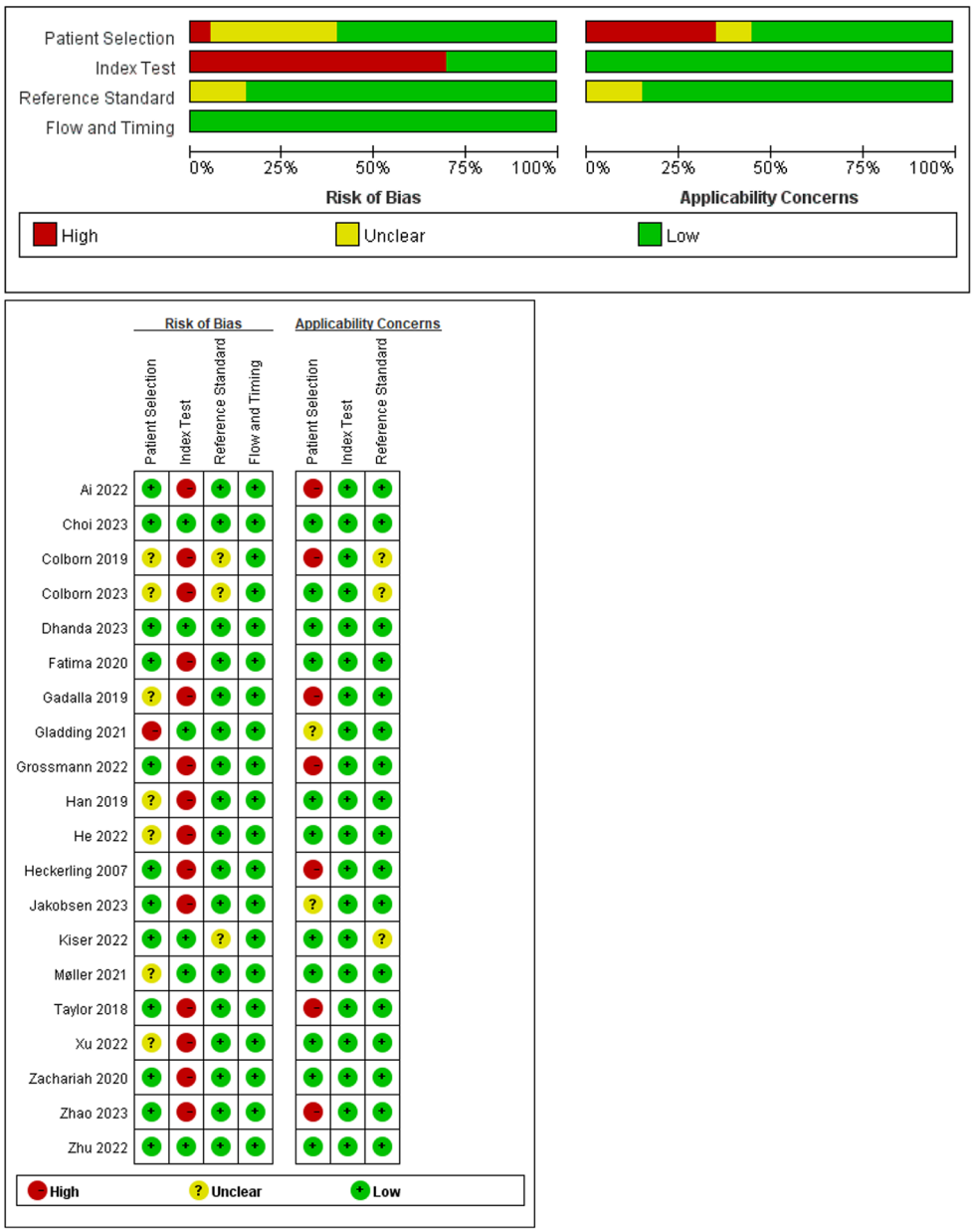

Fig. 2 Risk of bias and application concerns graph for the included studies

predictors' assigned weights in the final model and the results from multivariable analysis (77.77%). The summarized PROBAST results are presented in Supplementary Table S4. The results of Deeks' funnel plot (Fig. 3) suggested no significant publication bias ($P = 0.62$).

Results of meta-analysis

We extracted the best-performing model based on the AUC from each study for synthesis. Additionally, for studies that provided results from external validation, we also identified the model with the highest AUC on the external validation dataset. In total, 11 studies with 14 AI/ML models were included in the final meta-analysis [12–16, 19, 22, 25, 27, 29, 31]. The detailed characteristics of studies included in the meta-analysis, specific to each AI/ML algorithm, are presented in Table 1. The overall pooled AUC was 0.89 (95%CI 0.86–0.92). Additionally, the pooled Sen, Spe, PLR, NLR, and DOR were 0.78 (95%CI 0.71–0.84), 0.89 (95%CI 0.83–0.93), 6.99 (95%CI 4.38–11.14), 0.25 (95%CI 0.18–0.34) and 28.07 (95%CI 14.27–55.20), respectively. The forest plots for Sen and Spe are shown in Fig. 4, and the SROC curves are shown in Fig. 5.

The studies showed substantial heterogeneity, indicating significant variability among them ($I^2 > 99\%$). We investigated the sources of this heterogeneity through a threshold effect analysis, which showed no significant threshold effects, as indicated by the Spearman correlation coefficient

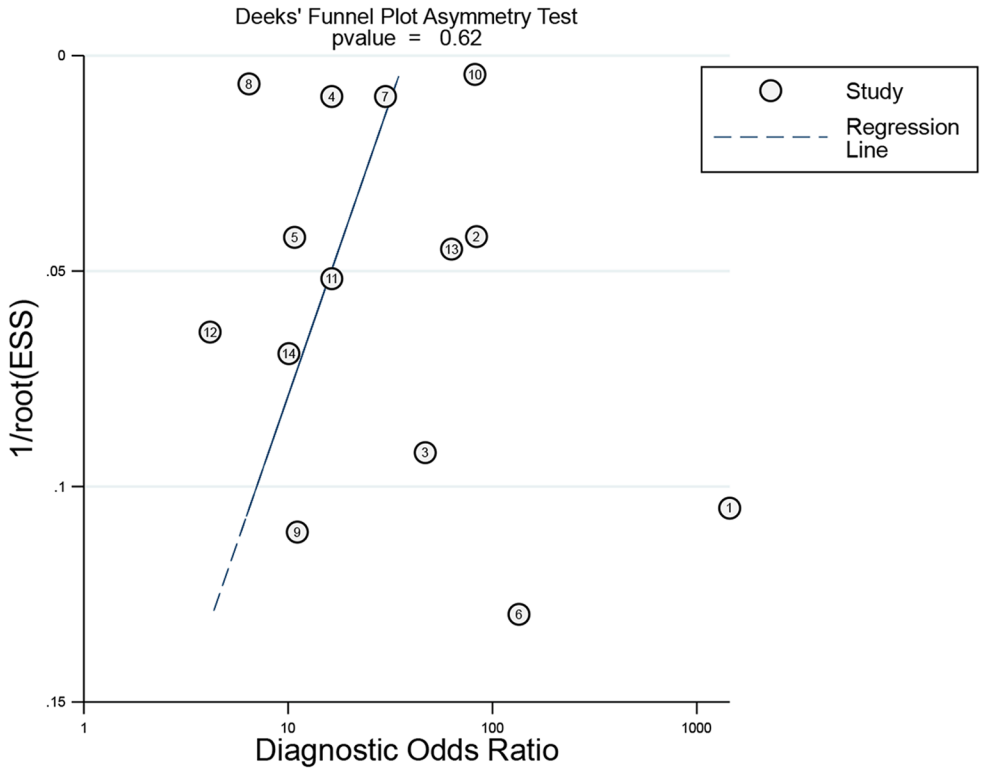

Fig. 3 Deeks' funnel plot to evaluate potential publication bias

(-0.262, $P=0.366$). Additionally, we performed a meta-regression to explore contributing factors to the observed heterogeneity, identifying participants, reference standard definition as influential factors (Fig. 6). The results of meta-regression indicate that the reference standard definition is likely associated with the observed heterogeneity ($P<0.01$). Due to the diagnostic gold standard for UTIs being a positive urine culture, some studies did not confirm urine culture results but instead used documented UTI diagnosis or ICD codes as the reference standard. This discrepancy in diagnostic criteria among studies may contribute to heterogeneity.

In order to investigate whether different validation dataset, study design, algorithm, participants, and definition exhibit heterogeneity in their combined results, we conducted subgroup analysis. Detailed results of the subgroup analysis are shown in Table 2. The slightly higher prediction accuracy on the internal validation dataset (AUC 0.91) compared to the external validation dataset (AUC 0.87) suggests some limitations in models' generalization capabilities when applied to new data. The minor decline in prediction accuracy observed in models employing a multi-center design (AUC 0.88) implies that collecting data from different locations may introduce more diversity and variability into the dataset. When stratified by AI/ML algorithm, models using RF algorithm demonstrate superior prediction accuracy (AUC 0.94). The predictive performance of models using the GB-based algorithm is also promising (AUC 0.91). Compared to models using urine culture as the reference standard definition, models using documented UTI diagnosis as the definition exhibit higher predictive performance (AUC 0.93). When employing "participants" as the grouping factor, the models utilizing inpatients' data showcases enhanced predictive capability (AUC 0.91). This could be because hospitalized patients' data may contain more clinical information or more laboratory tests data, enabling the models to better discern patterns relevant to UTI prediction.

Discussion

Model performance

Our combined results show predictive performance with a reasonably high AUC of 0.89, relatively strong specificity of 0.89, but moderate sensitivity of 0.78. This finding indicates that AI/ML algorithms have a low rate of misdiagnosis, but there is a significant probability of missed diagnosis. The decrease in sensitivity is mainly attributed to the class imbalance in datasets. In binary classification problems, if the target class is underrepresented, the AI/ML models may more readily predict the majority class, resulting in low sensitivity for the minority class. In reviewing the studies we included, the prevalence of UTI varied between 0.83% and 26.44%. Six of them even had a UTI prevalence of less than 2% [14, 15, 19, 25, 29, 31]. To address the issue of

Table 1 Summary of each individual AI/ML algorithm characteristics of studies included in the meta-analysis

Author	Year	Study design and setting	Type of validation set	Number of participants	UTI prevalence (%)	Best-performing AI/ML algorithm	AUC	Sensitivity	Specificity	TP	FP	FN	TN
Ai et al.	2022	Design: retrospective observational cohort study Single-center: Jingzhou Central Hospital Participants: surgery patients	internal	203	26 (12.81)	RF	0.918	0.962	0.983	25	3	1	174
Choi et al.	2023	Design: retrospective observational cohort study Multi-center: 2 tertiary hospitals Participants: surgery patients	external	62,255	18,903 (30.36)	XGBoost	0.967	0.855	0.933	16,162	2905	2741	40,447
Colborn et al.	2023	Design: retrospective observational cohort study Multi-center: 5 hospitals at University of Colorado Health system, ACS-NSQIP program Participants: surgery patients	internal	9189	144 (1.57)	LR	0.930	0.900	0.900	130	904	14	8141
Colborn et al.	2019	Design: retrospective observational cohort study Single-center: University of Colorado Hospital, ACS-NSQIP program Participants: surgery patients	internal	1646	30 (1.82)	RF	0.940	0.800	0.920	24	127	6	1489
Dhanda et al.	2023	Design: retrospective observational cohort study	internal	16,077	3566 (22.18)	NeedMicroXGB-Best	0.880	0.761	0.837	2714	2039	852	10,472
		Multi-center: 4 emergency departments and 1 outpatient family medicine department at the University of Kansas Medical Center Participants: emergency visits and outpatients	external	472	128 (27.12)	NoMicroRF-Best	0.850	0.789	0.814	101	64	27	280
Gladding et al.	2021	Design: retrospective observational nested cohort and case-control study Multi-center: Waitakere Hospital, North Shore Hospital Participants: inpatients and outpatients	external	14,514	61 (0.42)	RF	0.680	0.520	0.790	32	3035	29	11,418
He et al.	2022	Design: retrospective observational cohort study Multi-center: China-Japan Friendship Hospital, The First Affiliated Hospital of Zhengzhou University, The First Affiliated Hospital of China Medical University, Tianjin First Central Hospital Participants: inpatients	internal	822	180 (21.90)	GBDT	0.831	0.594	0.880	107	77	73	565

Table 1 (continued)

Author	Year	Study design and setting	Type of validation set	Number of participants	UTI prevalence (%)	Best-performing AI/ML algorithm	AUC	Sensitivity	Specificity	TP	FP	FN	TN
Kiser et al.	2022	Design: retrospective observational cohort study	internal	1733	15 (0.87)	LR	0.936	0.868	0.954	13	79	2	1639
		Multi-center: University of Utah Health, Intermountain Healthcare, ACS-NSQIP program Participants: surgery patients	external	15,434	125 (0.81)	LR	0.929	0.712	0.962	89	577	36	14,732
Taylor et al.	2018	Design: retrospective observational cohort study Single-center: 4 emergency departments at a health care system Participants: emergency visits	internal	16,077	3566 (22.18)	XGBoost	0.904	0.617	0.949	2200	638	1366	11,873
Zacha-riah et al.	2020	Design: retrospective observational cohort study Multi-center: 3 acute care hospitals at a large university hospital system in New York Participants: inpatients	internal	358,938	5904 (1.64)	DT	0.780	0.782	0.642	4617	126,386	1287	226,648
Zhu et al.	2022	Design: retrospective observational cohort study	internal	797	21 (2.63)	Ensemble	0.821	0.809	0.723	17	215	4	561
		Multi-center: 25 hospitals participated in CCBPC program Participants: inpatients	external	3837	53 (1.38)	Ensemble	0.808	0.811	0.701	43	1131	10	2653

Note: RF = random forest, XGBoost = extreme gradient boosting, LR = logistic regression, GBDT = gradient boosting decision tree, DT = decision tree, ACS-NSQIP = American College of Surgeons National Surgical Quality Improvement Program, CCBPC = Common Complications of Bedridden Patients and the Construction of Standardized Nursing Intervention Model

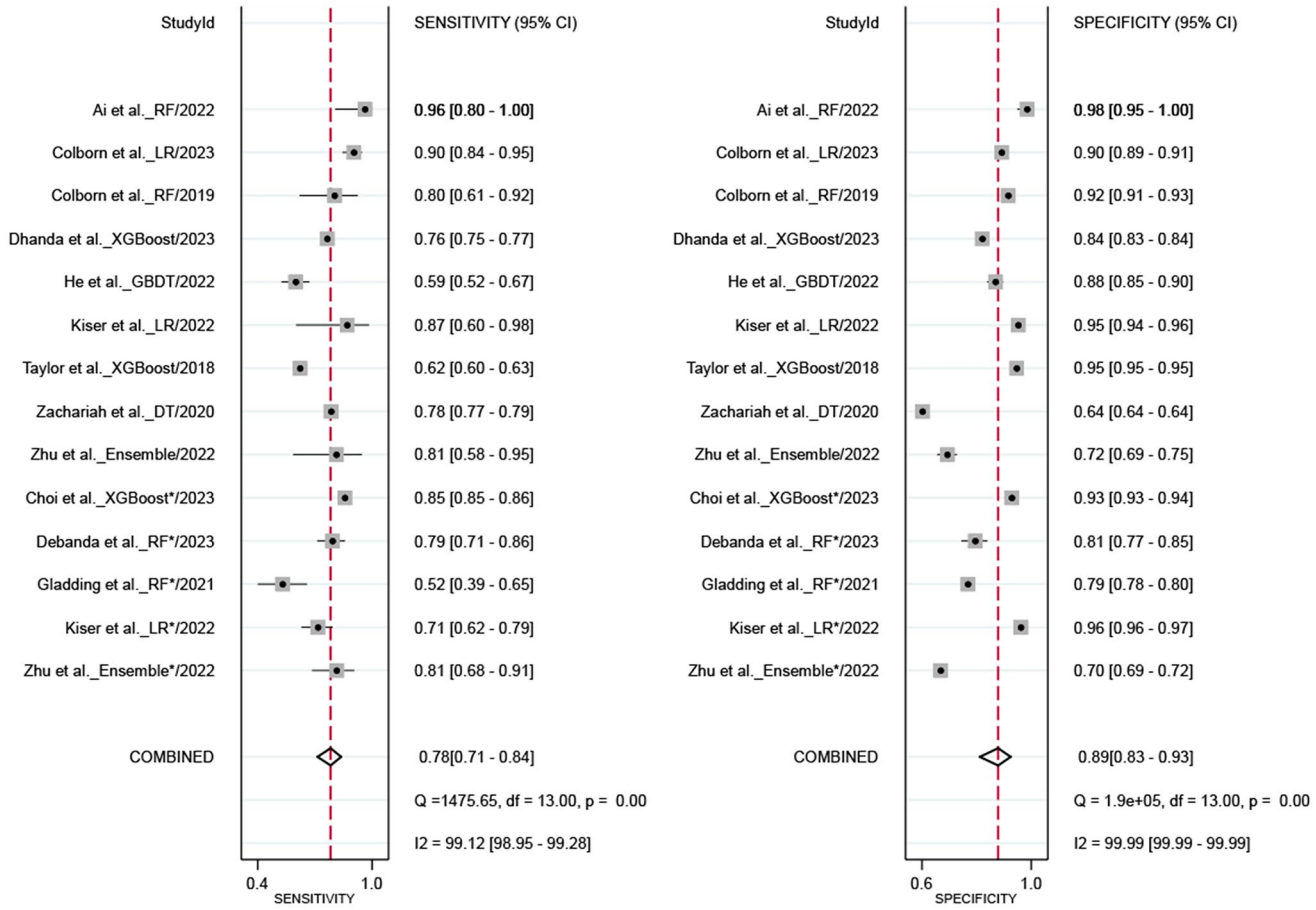

Fig. 4 Forest plot for the pooled sensitivity and specificity estimates

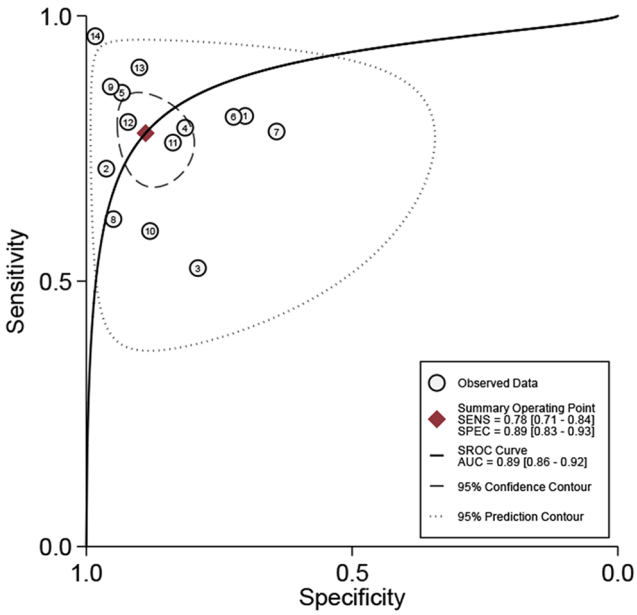

Fig. 5 SROC curve of AI/ML algorithms for the prediction of UTIs

imbalanced datasets, resampling and feature selection are commonly used methods. The application of resampling can enhance the model's learning ability regarding the minority class [22, 25, 29, 31]. Additionally, carefully selected variables can make the model more focused on crucial information and improve the recognition rate of the minority class [15]. Considering that both resampling and feature selection carry a risk of overfitting, it is suitable to combine them with cross-validation procedure to assess model performance. Meanwhile, using ensemble learning methods can also improve overall performance in dealing with class imbalance [32]. Therefore, more optimized AI/ML models are demanded to achieve accurate detection of UTI.

Heterogeneity

We conducted meta-regression to identify the sources of heterogeneity, followed by subgroup analysis. Substantial heterogeneity was observed in definition ($P < 0.01$). As we know, UTI diagnostic criteria can vary not only among different studies but also across various healthcare settings and geographic regions globally. However, despite these

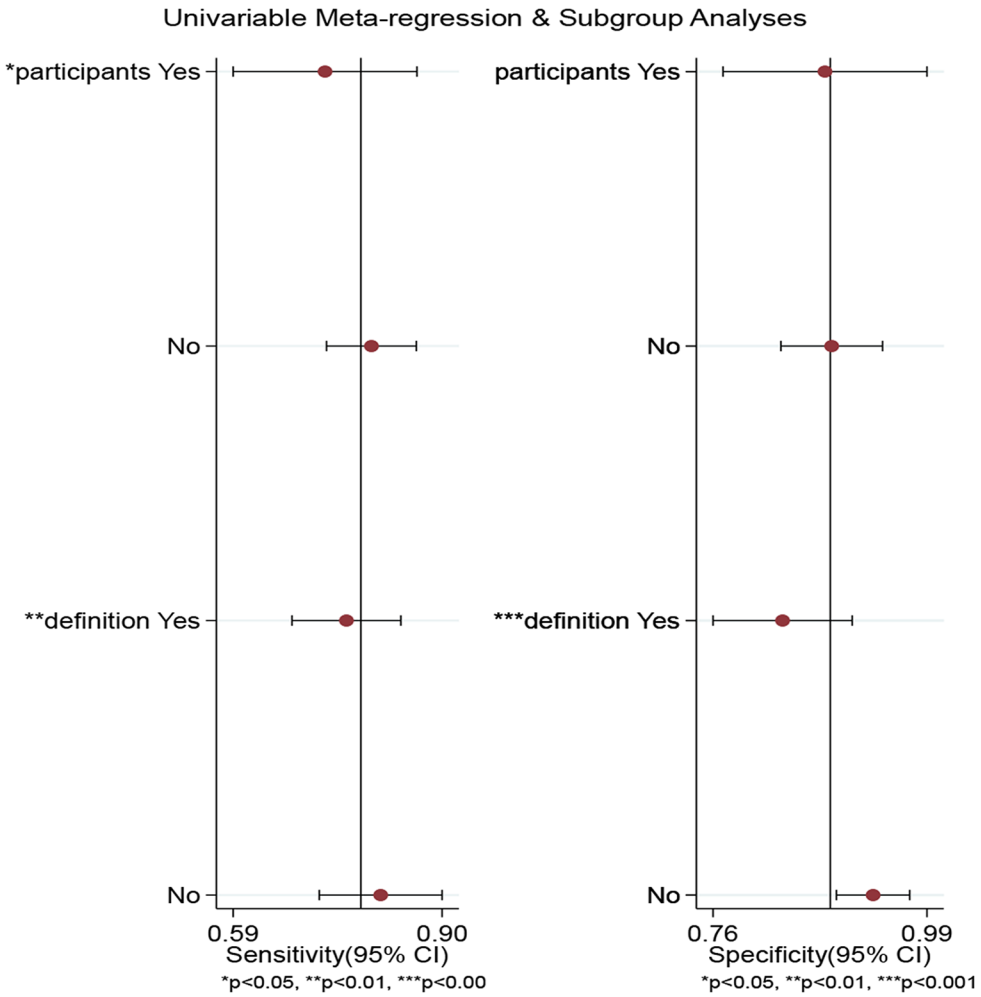

Fig. 6 Univariate meta-regression plot of all AI/ML models

variations, there is generally a consensus that UTI diagnosis revolves around the presence of bacterial growth in urine cultures (Supplementary Table S5). Considering the limited number of studies included, we did not further subdivide the urine culture category based on specific colony-forming unit (CFU) criteria, such as $>10^5$, $>10^4$, and $>10^3$, for more detailed analysis. The decline in predictive accuracy observed within the urine culture definition group stemmed from the lack of uniformity in the criteria for defining positive CFU counts in urine cultures. This variability consequently introduced a confounding effect on the combined results. For example, if a standard of $>10^5$ is used, specimens with counts $>10^3$ but $\leq 10^5$ would be deemed negative. However, in a model with a standard of $>10^3$, these specimens would be considered positive. This inconsistency raises questions about the accuracy of predictive results. In contrast, the models that used a definition of documented UTI diagnosis showed superior predictive accuracy, likely because most models used data sourced from the ACS-NQSIP database [14, 15, 25], where UTI diagnostic standard is consistent. This consistency in UTI diagnosis within the same database helps reduce heterogeneity. As far as we are concerned, further research and interpretation are needed when fitting models to data obtained from different diagnostic criteria.

Clinical utility

Traditionally, clinicians rely on symptoms, general examinations, and lab tests to "infer" whether a patient has a UTI before urine culture results are available. When patients present with classical UTI symptoms, including dysuria, urinary frequency, urgency, and hematuria, the sensitivity based on these typical UTI symptoms is between 50% and 80% [33, 34]. However, when patients present with atypical symptoms, especially those unable to articulate their condition, such as ICU patients, AI/ML models can serve as excellent diagnostic aids. These models leverage a broader array of data, much like the predictors used in the included studies, encompassing demographics, symptoms, laboratory data from blood and urine samples, comorbidities, surgery information, disease treatment and clinical course,

Table 2 Subgroup analysis of the performance of AI/ML application in UTI identification and prediction

Subgroup	Number of models	AUC	Sensitivity	Specificity	PLR	DLR	DOR
All combined	14	0.89 (0.86–0.92)	0.78 (0.71–0.84)	0.89 (0.83–0.93)	6.99 (4.38–11.14)	0.25 (0.18–0.34)	28.07 (14.27–55.20)
Validation dataset							
internal	9	0.91 (0.88–0.93)	0.80 (0.70–0.87)	0.90 (0.82–0.94)	7.79 (4.23–14.36)	0.23 (0.15–0.35)	34.41 (13.71–86.38)
external	5	0.87 (0.84–0.90)	0.76 (0.65–0.84)	0.87 (0.76–0.94)	5.86 (2.88–11.93)	0.28 (0.18–0.42)	20.92 (7.68-57.00)
Design							
multi-center	8	0.88 (0.85–0.90)	0.77 (0.71–0.83)	0.86 (0.79–0.91)	5.52(3.51–8.66)	0.26 (0.19–0.35)	21.02 (10.87–40.64)
Algorithm							
RF	4	0.94 (0.91–0.95)	0.82 (0.59–0.94)	0.91 (0.78–0.96)	8.84 (2.98–26.22)	0.20 (0.07–0.56)	45.31 (5.61-366.03)
GB-based	4	0.91 (0.88–0.93)	0.72 (0.60–0.82)	0.91 (0.86–0.94)	7.94 (4.91–12.85)	0.30 (0.20–0.46)	25.06 (12.38–54.86)
Definition							
urine culture	8	0.85 (0.82–0.88)	0.76 (0.69–0.82)	0.84 (0.75–0.90)	4.66 (2.99–7.25)	0.29 (0.22–0.37)	16.25 (9.34–28.28)
documented UTI diagnosis	5	0.93 (0.91–0.95)	0.78 (0.63–0.88)	0.92 (0.86–0.95)	9.84 (5.26–18.41)	0.24 (0.13–0.43)	41.36 (13.76-124.33)
Participants							
inpatients	10	0.91(0.88–0.93)	0.82 (0.74–0.87)	0.90 (0.82–0.94)	8.01 (4.34–14.77)	0.20 (0.14–0.30)	39.10 (16.42–93.13)

Note: RF = random forest; GB-based = algorithms using gradient boosting

ICD-9 codes and ICD-10 codes, and past medical history. By integrating such comprehensive datasets, AI/ML models can enhance diagnostic accuracy, particularly in cases where typical symptoms are absent or challenging to assess. This multifaceted approach enables clinicians to make more informed decisions, especially in complex clinical scenarios, ultimately improving patient outcomes. The clinical utility of the AI/ML models extends beyond diagnosis to infection monitoring. The current HAI monitoring systems have infections alert functions, but they mainly rely on capturing abnormal body temperatures, abnormal laboratory reports, or infection-related keywords in medical records. This monitoring mode is inefficient, and the accuracy of infection alerts is questionable. However, envisioning the future, if we could integrate AI/ML models for predicting UTI or other HAI into monitoring systems, we would be able to prospectively predict HAI using AI/ML algorithms. This would be of significant importance in preventing nosocomial infection outbreaks and controlling multidrug-resistant organisms.

Strengths and limitations

To our knowledge, our systematic review and meta-analysis is the first quantitative evaluation of AI/ML algorithms' diagnostic accuracy and reliability for UTI. However, there are a couple of limitations that need to be aware of. Firstly, the number of studies included in the meta-analysis is relatively small and insufficient. Given the rapid advancements in this field, our research may have missed some studies that could have been included. Secondly, considerable heterogeneity was found in our meta-analysis. The reference standard definition may be the main source of heterogeneity. Thirdly, the diversity of predictors may affect the models' diagnostic performance of UTI. We did not conduct relevant subgroup analysis due to lack of detailed information. Additionally, we reconstructed the 2×2 confusion matrix from included studies without specific TP, FP, FN, or TN numbers, which might introduce misclassification due to rounding errors.

With larger datasets comprising more patients and variables, AI/ML may become a very powerful and accurate tool for prediction of UTIs. Utilizing AI/ML algorithms can assist clinicians in detecting and identifying patients at high risk of UTIs, enabling timely treatment and management. Future research with advanced AI/ML algorithms, large-scale longitudinal cohorts, and enhanced clinical interpretability is necessary to improve predictive performance for UTIs.

Conclusion

In conclusion, this systematic review and meta-analysis quantitatively evaluated the application of AI/ML algorithms in identification and prediction in UTI for the first time. Our results indicate that AI/ML models provide promising performance for UTI identification and prediction. The use of AI/ML algorithms can assist clinicians detect and identify patients at high risk of UTIs, enabling timely treatment and management. It is hoped that with further refinement and validation, AI/ML algorithm-based prediction models can be gradually integrated into clinical practice.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00345-024-05145-4.

Acknowledgements Not applicable.

Author contributions LS carried out the study and drafted the manuscript; YG conceived of the study, participated in its design and coordination. LS, JA, NW and JY conducted literature research, study selection, quality assessment and data extraction; LS and JA conducted the statistical analysis; LS, JW and YG contributed to data interpretation and revised the manuscript. All authors read and approved the final manuscript.

Funding Project supported by Natural Science Basic Research Planning Program of Shaanxi Province, China (Grant No.2024JC-YBMS-607).

Declarations

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Competing interests The authors declare that they have no competing interests.

References

1. Foxman B (2010) The epidemiology of urinary tract infection. Nat Reviews Urol 7(12):653–660
2. Gupta K, Hooton TM, Roberts PL et al (2001) Patient-initiated treatment of uncomplicated recurrent urinary tract infections in young women. Ann Intern Med 135(1):9–16
3. Flores-Mireles A, Hreha TN, Hunstad DA (2019) Pathophysiology, treatment, and Prevention of Catheter-Associated urinary tract infection. Top Spinal cord Injury Rehabilitation 25(3):228–240
4. Saint S (2000) Clinical and economic consequences of nosocomial catheter-related bacteriuria. Am J Infect Control 28(1):68–75
5. Wang J, Liu F, Tartari E et al (2018) The prevalence of Healthcare-Associated infections in Mainland China: a systematic review and Meta-analysis. Infect Control Hosp Epidemiol 39(6):701–709
6. Zhu C, Zi H, Huang Q et al (2021) Analysis of the disease burden of urinary tract infections in China from 1990 to 2019. J Mod Urol 26(5):376–381
7. Ross J, Hickling D (2022) Medical treatment for urinary tract infections. Urologic Clin North Am 49(2):283–297
8. Hecker MT, Fox CJ, Son AH et al (2014) Effect of a stewardship intervention on adherence to uncomplicated cystitis and pyelonephritis guidelines in an emergency department setting. PLoS ONE 9(2):e87899
9. Goździkiewicz N, Zwolińska D, Polak-Jonkisz D (2022) The Use of Artificial Intelligence algorithms in the diagnosis of urinary tract Infections-A literature review. J Clin Med. ;11(10)
10. Whiting PF, Rutjes AW, Westwood ME et al (2011) QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 155(8):529–536
11. Moons KGM, Wolff RF, Riley RD et al (2019) PROBAST: A Tool to assess risk of Bias and Applicability of Prediction Model studies: explanation and elaboration. Ann Intern Med 170(1):W1–w33
12. Ai J, Hu Y, Zhou FF et al (2022) Machine learning-assisted ensemble analysis for the prediction of urinary tract infection in elderly patients with ovarian cancer after cytoreductive surgery. World J Clin Oncol 13(12):967–979
13. Choi MH, Kim D, Park Y et al (2024) Development and validation of artificial intelligence models to predict urinary tract infections and secondary bloodstream infections in adult patients. J Infect Public Health 17(1):10–17
14. Colborn KL, Bronsert M, Hammermeister K et al (2019) Identification of urinary tract infections using electronic health record data. Am J Infect Control 47(4):371–375
15. Colborn KL, Zhuang Y, Dyas AR et al (2023) Development and validation of models for detection of postoperative infections using structured electronic health records data and machine learning. Surgery 173(2):464–471
16. Dhanda G, Asham M, Shanks D et al (2023) Adaptation and external validation of pathogenic urine culture prediction in primary care using machine learning. Ann Fam Med 21(1):11–18
17. Fatima N, Zheng H, Massaad E et al (2020) Development and Validation of Machine Learning Algorithms for Predicting adverse events after surgery for lumbar degenerative spondylolisthesis. World Neurosurg 140:627–641
18. Gadalla AAH, Friberg IM, Kift-Morgan A et al (2019) Identification of clinical and urine biomarkers for uncomplicated urinary tract infection using machine learning algorithms. Sci Rep 9(1):19694
19. Gladding PA, Ayar Z, Smith K et al (2021) A machine learning PROGRAM to identify COVID-19 and other diseases from hematology data. Future Sci OA 7(7):Fso733
20. Grossmann NC, Schuettfort VM, Betschart J et al (2022) Risk factors for concomitant positive midstream urine culture in patients presenting with symptomatic ureterolithiasis. Urolithiasis 50(3):293–302
21. Han SS, Azad TD, Suarez PA et al (2019) A machine learning approach for predictive models of adverse events following spine surgery. Spine J 19(11):1772–1781
22. He Y, Peng P, Ying W et al (2022) Contrast between traditional and machine learning algorithms based on a urine culture predictive model: a multicenter retrospective study in patients with urinary calculi. Transl Androl Urol 11(2):139–148
23. Heckerling PS, Canaris GJ, Flach SD et al (2007) Predictors of urinary tract infection based on artificial neural networks and genetic algorithms. Int J Med Informatics 76(4):289–296
24. Jakobsen RS, Nielsen TD, Leutscher P et al (2023) Clinical explainable machine learning models for early identification of patients at risk of hospital-acquired urinary tract infection. J Hosp Infect
25. Kiser AC, Eilbeck K, Ferraro JP et al (2022) Standard vocabularies to Improve Machine Learning Model Transferability with Electronic Health Record Data: Retrospective Cohort Study using Health Care-Associated infection. JMIR Med Inf 10(8):e39057

26. Møller JK, Sørensen M, Hardahl C (2021) Prediction of risk of acquiring urinary tract infection during hospital stay based on machine-learning: a retrospective cohort study. PLoS ONE 16(3):e0248636
27. Taylor RA, Moore CL, Cheung KH et al (2018) Predicting urinary tract infections in the emergency department with machine learning. PLoS ONE 13(3):e0194085
28. Xu Z, Zhu C, Gu Y et al (2022) Developing a siamese network for UTIs risk prediction in Immobile patients undergoing stroke. Stud Health Technol Inf 290:714–718
29. Zachariah P, Sanabria E, Liu J et al (2020) Novel strategies for Predicting Healthcare-Associated infections at Admission: implications for nursing care. Nurs Res 69(5):399–403
30. Zhao YJ, Chen CY, Huang ZY et al (2023) Prediction of upcoming urinary tract infection after intracerebral hemorrhage: a machine learning approach based on statistics collected at multiple time points. Front Neurol. ;14
31. Zhu C, Xu Z, Gu Y et al (2022) Prediction of post-stroke urinary tract infection risk in immobile patients using machine learning: an observational cohort study. J Hosp Infect 122:96–107
32. Galar M, Fernandez A, Barrenechea E et al (2012) A review on ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybernetics Part C (Applications Reviews) 42(4):463–484
33. Giesen LG, Cousins G, Dimitrov BD et al (2010) Predicting acute uncomplicated urinary tract infection in women: a systematic review of the diagnostic accuracy of symptoms and signs. BMC Fam Pract 11:78
34. Schmiemann G, Kniehl E, Gebhardt K et al (2010) The diagnosis of urinary tract infection: a systematic review. Deutsches Arzteblatt Int 107(21):361–367