Clin. Lab. 2026;72:XXX-XXX
©Copyright

ORIGINAL ARTICLE

A Machine Learning Framework for Predicting Nosocomial *Escherichia coli* Infections in Cervical Cancer

Wulin Shan [1, 2, *], Qingqing Shan [3, *], Xinxin Xu [4], Jun Chen [5], Wenju Peng [6, 7], Ming Li [1, 2]

[*] *These authors contributed equally to this work and share the first authorship*
[1] *Department of Laboratory Diagnostics, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China*
[2] *Department of Laboratory Diagnostics, Anhui Provincial Cancer Hospital, Hefei, China*
[3] *Clinical Pathology Center, The First Affiliated Hospital of Anhui Medical University, Anhui Public Health Clinical Center, Hefei, PR China*
[4] *Laboratory Diagnostics, Anhui Institute of Medicine, Hefei, PR China*
[5] *Laboratory Diagnostics, Bozhou University, Hefei, PR China*
[6] *Department of Obstetrics and Gynecology, The First Affiliated Hospital of USTC, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China*
[7] *Department of Gynecologic Oncology, Anhui Provincial Cancer Hospital, Hefei, China*

ABSTRACT

Background: **This study aimed to characterize the etiological profile of nosocomial infections in cervical cancer patients and to develop a machine learning-based prediction model for infections caused by the predominant pathogen, *Escherichia coli*, to support clinical decision-making in anti-infective therapy and risk stratification.**
Methods: **We conducted a retrospective analysis of clinical data from 118 cervical cancer patients to evaluate the distribution and antimicrobial resistance patterns of infectious pathogens. Predictive factors for *Escherichia coli* infection were identified, and a corresponding prediction model was developed. All the statistical analyses were carried out via R software (version 4.3.2) and iResearch (version 2.9.2).**
Results: **A total of 151 pathogenic isolates were obtained, with the highest prevalence detected in mid-stream urine samples (69.54%, 105/151). Gram-negative bacteria constituted 76.82% (116/151) of the isolates, among which *Escherichia coli* was the most frequently identified species (50.33%, 76/151). Antimicrobial susceptibility testing revealed resistance rates exceeding 55% to ceftriaxone, ciprofloxacin, trimethoprim-sulfamethoxazole, and levofloxacin among *Escherichia coli* isolates, whereas high susceptibility was retained to carbapenems, piperacillin-tazobactam, and amikacin. Logistic regression analysis revealed that *Escherichia coli* infection was positively associated with earlier clinical stage, absence of anemia, and mid-stream urine sample type. Within the urinary infection subgroup, positive urinary nitrite was also correlated with increased infection risk. Feature selection utilizing multiple approaches informed the construction of the prediction model.**
Logistic regression and svm_cross_validation exhibited stable performance in the full sample analysis. Restricting the analysis to mid-stream urine samples substantially improved model performance. The svm-based model yielded AUC values of 0.81 and 0.89 in the training and test sets, respectively, and the logistic model achieved AUCs of 0.87 and 0.90, respectively.
Conclusions: **Nosocomial infections in cervical cancer patients are caused primarily by gram-negative bacilli within the urinary tract, with *Escherichia coli* representing the most prevalent pathogen. The machine learning model, which incorporates readily available clinical parameters such as disease stage, anemia status, and urinalysis results, demonstrated robust discriminatory performance in predicting *Escherichia coli* infection in mid-stream urine samples. This tool offers a practical approach for risk identification and guides a more targeted empiric therapy, holding promise for improving treatment outcomes in patients with cervical cancer.**
(Clin. Lab. 2026;72:xx-xx. DOI: 10.7754/Clin.Lab.2025.250926)

Manuscript accepted November 25, 2025

Correspondence:
Ming Li
Department of Laboratory Diagnostics
The First Affiliated Hospital of USTC
Division of Life Sciences and Medicine
University of Science and Technology of China
Hefei, 230031
China
Email: lm831216@ustc.edu.cn

Wenju Peng
Department of Gynecologic Oncology
Anhui Provincial Cancer Hospital
Huanhudong Road 107
Hefei, 230031
China
Phone/Fax number: + 86 0551-65327608
Email: pwenju@foxmail.com

KEYWORDS

cervical cancer, machine learning, pathogenic bacteria, nosocomial infection, prediction model

INTRODUCTION

Cervical cancer represents one of the most prevalent malignancies affecting women globally [1,2]. Treatments such as surgery, radiotherapy, and chemotherapy frequently result in immunosuppression, significantly increasing the risk of nosocomial infections among these patients [1,3]. These infections can lead to prolonged hospitalization, elevated healthcare costs, and potentially life-threatening complications including sepsis, thereby adversely affecting clinical outcomes. The growing prevalence of antimicrobial resistance driven in part by inappropriate antibiotic use has further complicated treatment strategies, with notably high resistance rates to fluoroquinolones and third-generation cephalosporins increasingly reported [4,5].

Current detection approaches for infections in cervical cancer patients rely heavily on microbial culture and antibiotic susceptibility testing. However, these methods are time intensive and offer limited utility for early clinical intervention [6,7]. Although conventional biomarkers such as procalcitonin and C-reactive protein are widely used, they lack specificity for pathogenic organisms, reducing their diagnostic precision. Furthermore, traditional statistical models often prove inadequate when applied to complex, multidimensional clinical data.

Machine learning algorithms have emerged as powerful tools in predictive healthcare, demonstrating considerable success in the early detection of conditions such as sepsis and ventilator-associated pneumonia [8,9]. Despite these advances, the application of such techniques to predict infections specifically in cervical cancer patients remains limited. Consequently, a thorough investigation of pathogen distribution and resistance patterns, complemented by the development of machine learning models capable of integrating diverse clinical variables, is essential to facilitate early diagnosis and promote rational antimicrobial use, ultimately improving patient care and survival.

In this study, we retrospectively analyzed clinical and microbiological data from cervical cancer patients to delineate the profiles and resistance characteristics of pathogenic infections. Using machine learning algorithms, including logistic regression and support vector machine (svm), we developed a predictive model for *Escherichia coli* (*E. coli*) infections that exhibited strong discriminatory power, with the logistic regression model achieving an area under the receiver operating characteristic curve (AUC) of 0.90 in the test set. By innovatively leveraging machine learning for the prediction of *E. coli* infections among culture-positive cases, this work establishes a framework for personalized infection management and targeted antibiotic therapy, with the potential to enhance treatment efficacy.

MATERIALS AND METHODS

Study setting and population

This retrospective cohort study was conducted at the West District of the First Affiliated Hospital of University of Science and Technology of China (Anhui Provincial Cancer Hospital), the only tertiary-care specialized cancer hospital in Anhui Province. The facility maintains 1,696 inpatient beds and 29 clinical subspecialties. Its gynecologic oncology unit is a provincial-level key specialty, serving a broad regional population, which ensures sufficient patient enrollment. A structured infection control program is implemented hospital-wide, with prospective audits and feedback from clinicians and pharmacists providing timely insight into local pathogen distribution and antimicrobial resistance patterns, thereby promoting rational antimicrobial use.

This study included consecutive 118 hospitalized cervical cancer patients with positive microbial cultures. For urine cultures, a urinary tract infection was diagnosed in cases with colony counts meeting the threshold of > 10^5 CFU/mL for gram-negative bacilli, > 10^4 CFU/mL for gram-positive bacteria, or > 10^4 CFU/mL for fungi from a mid-stream urine specimen. In the presence of significant pyuria (> 1 + WBC on qualitative urinalysis), a bacterial colony count exceeding 10^3 CFU/mL was also considered diagnostic for a urinary tract infection. Polymicrobial cultures (growth of ≥ 3 potential pathogens) were excluded from the predictive modeling analysis to ensure a clear outcome label.

Clinical data collection

Demographic and clinical variables, including age, disease stage, geographic region, and standardized laboratory assessments, were retrospectively collected from medical records. These included the biochemical parameters glucose (GLU), Lpa, ALT and AST for hepatic function, creatinine (CREA) and blood urea nitrogen (BUN) for renal function, the tumor biomarker squamous cell carcinoma antigen (SCC), complete blood count indices - WBC, lymphocyte (L) and neutrophil (N) counts and percentages for infection; RBC and hemoglobin (Hb) for anemia; platelet (PLT) count for coagulation, as well as the qualitative urinalysis measures: occult blood (U_OB), protein (U_PRO), white blood cell (U_WBC1) and nitrite (U_nitrite), and the quantitative parameters: white blood cell (U_WBC). A rigorous data integrity protocol was implemented, featuring dual independent extraction and entry followed by cross-verification by a third investigator.

Pathogen isolation, culture, and antimicrobial susceptibility

The test specimen types included deep sputum, fresh whole blood, mid-stream urine, purulent secretions, vaginal secretions, drainage fluid, and stool. Isolation and culture procedures were performed in accordance with the National Clinical Laboratory Operating Procedures (4th Edition). Bacterial identification and antimicrobial susceptibility testing were conducted via the VITEK-2 compact automated microbial analysis system (bioMérieux, France). The minimum inhibitory concentration (MIC) dilution method was employed for susceptibility testing, with results interpreted based on the Clinical and Laboratory Standards Institute 2020 guidelines (CLSI M100-30). The quality control strains included *E. coli* (ATCC 25922) and *Staphylococcus aureus* (ATCC 25923).

Construction of infection prediction models

For analyses incorporating all specimen types, predictor variables, including specimen type, disease stage, and complete blood count parameters (WBC, RBC, and Hb) were selected based on logistic univariate regression results ($p < 0.1$) and decision tree outcomes. For analyses restricted to mid-stream urine specimens, variables were selected via logistic univariate regression ($p < 0.1$), decision tree, and least absolute shrinkage and selection operator (LASSO), encompassing disease stage, complete blood count parameters (WBC, RBC, Hb), and urinalysis indicators (WBC and nitrite). All above analyses were performed via R software (version 4.3.2). We employed the *forest plot* package for regression analysis and the *glmnet* package for feature selection via LASSO regression. Model performance was evaluated by generating receiver ROC curves using the *pROC* package. All data visualizations were created with ggplot2. The following machine learning algorithms analyzed by iResearch (version 2.9.2) were applied to evaluate their utility in infection prediction: Naive Bayes classifier, K-nearest neighbors (KNN) classifier, Logistic regression classifier, Random forest classifier, Decision tree classifier, Artificial neural network (ANN) classifier, Support vector machine with cross-validation (svm_cross_validation), Gradient boosting classifier (ensemble learning), LightGBM, Adaptive boosting (AdaBoost) classifier and XGBoost classifier.

Statistical analysis

Data processing and statistical analyses were conducted using R version 4.3.2 and iResearch (version 2.9.2). Continuous variables are expressed as the mean ± standard deviation ($x \pm s$), and categorical variables are expressed as frequencies (%). Univariate logistic regression was used to assess associations between clinical characteristics/laboratory indicators and pathogenic infections. Variables with $p < 0.1$ were included in model construction, and $p < 0.05$ was considered statistically significant.

RESULTS

Clinical characteristics and pathogen distribution

A cohort of 118 cervical cancer patients with postoperative infections were enrolled, with ages ranging from 24 to 73 years. Among these, 74 patients (62.71%) were over 50 years of age. The distribution according to the International Federation of Gynecology and Obstetrics (FIGO) staging system was as follows: 37 patients were in stage I, 25 were in stage II, 38 were in stage III, and 18 were in stage IV. Geographically, 42 patients were from southern Anhui, and 71 were from northern Anhui (Table 1).

Microbiological analysis identified 151 pathogenic isolates from all the samples. Gram-negative bacilli predominated, comprising 116 strains (76.82%), with *E. coli* being the most prevalent pathogen (76 strains, 50.33%). A total of 29 gram-positive (19.21%) were identified, primarily *Staphylococcus* and *Enterococcus* species, along with six fungal isolates (3.97%), all belonging to the *Candida genus*. Mid-stream urine samples were the most frequent source of pathogens, accounting for 105 isolates (69.54%), including 61 strains of *E. coli*. Other sources included fresh whole blood (15 strains, 9.93%) and deep sputum (9 strains, 5.96%). Notably, seven *Staphylococcus* isolates were recovered from blood samples, six of which were coagulase-negative species (*Staphylococcus epidermidis* and *Staphylococcus hominis*), underscoring the importance of stringent sampling protocols to minimize contamination (Table 2).

Antimicrobial resistance profiles of major pathogens

Antimicrobial resistance was assessed for the most frequently isolated gram-negative and gram-positive pathogens. *E. coli* exhibited high resistance to ceftriaxone and ciprofloxacin (both > 65%), as well as to trimethoprim-sulfamethoxazole and levofloxacin (> 55%). In contrast, resistance rates to nitrofurantoin, imipenem, piperacillin/tazobactam, and amikacin were less than 10%. *Proteus mirabilis* showed elevated resistance (> 50%) to nitrofurantoin, trimethoprim-sulfamethoxazole, and ciprofloxacin. Similarly, *Klebsiella pneumoniae* was highly resistant to trimethoprim-sulfamethox-

Table 1. Clinical characteristics of cervical cancer patients with bacterial infections.

Clinical characteristics	Number (n = 118)	Proportion (%)
Age		
≤ 50	44	37.29
> 50	74	62.71
FIGO stage		
I	37	31.36
II	25	21.19
III	38	32.20
IV	18	15.25
Place of residence		
Northern Anhui region	42	35.59
Southern Anhui region	71	60.17
Others	5	4.24

In the staging category, "Other" represents cases with no definitive staging. In the place of residence category, "Other" represents patients not from Anhui Province.

azole, ciprofloxacin, and ceftriaxone. No resistance to carbapenems, piperacillin/tazobactam, cefepime, or amikacin was detected in the latter two species. Among gram-positive bacteria, *Enterococcus faecalis* exhibited high resistance (> 65%) to quinupristin/dalfopristin, tetracycline, and erythromycin. *Staphylococcus aureus* presented the highest resistance rate to penicillin, whereas *Staphylococcus hominis* was frequently resistant to penicillin and erythromycin. All gram-positive cocci remained fully susceptible to vancomycin, linezolid, and tigecycline (Tables 3 and 4).

Development and validation of the prediction model using all sample types

In the primary analysis encompassing all sample types, cultures positive for *E. coli* were designated as the positive cohort, with samples containing other pathogens constituted the negative cohort for subsequent bioinformatic investigation. Univariate logistic regression revealed significant positive associations between *E. coli* infection and earlier clinical stage (p = 0.008) and midstream urine sample type (p = 0.012), and negative correlations with anemia-related indicators (RBC: p = 0.087; Hb: p = 0.083; Figure 1A). No statistically significant differences were observed in hepatic (ALT: p = 0.911, AST: p = 0.475) or renal function (CREA: p = 0.507, BUN: p = 0.962) parameters between the compared groups, indicating that impaired liver or kidney function is unlikely to constitute a primary risk factor for *E. coli* infection. Decision tree-based feature selection highlighted blood WBC and clinical stage as key predictors (Figure 1B). Consequently, disease stage, RBC, WBC, Hb, and sample type were used for predictive modeling. Pearson’s correlation analysis confirmed the absence of multicollinearity (all coefficients < 0.5; Figure 1C).

Eleven machine learning algorithms were evaluated. The naive_bayes, logistic, and svm_cross models showed relatively stable performance, whereas the other models indicated potential overfitting (Figure 2A and 2B). However, the discriminatory capacity was limited, with all the AUC values being less than 0.75. Subsequent feature combination analyses did not yield improvements, with no model achieving an AUC above 0.7 (Figure 2C and 2D). These results indicate that considerable optimization is essential for translating the preliminary potential of these models into reliable performance.

Development and validation of a prediction model using mid-stream urine samples

For the focused analysis of midstream urine samples, we employed a consistent case definition, designating *E. coli*-positive cultures as the case group and those containing other pathogens as the control group for subsequent bioinformatic investigation. Univariate logistic regression confirmed significant positive associations between *E. coli* infection and earlier clinical stage (p = 0.001) and positive urinary nitrite (p = 0.013), and negative correlations with anemia indicators (RBC: p = 0.018; Hb: p = 0.093; Figure 3A). No statistically significant differences were observed in hepatic (ALT: p = 0.726, AST: p = 0.942) or renal function (CREA: p = 0.214, BUN: p = 0.969) parameters between the compared groups, indicating that impaired liver or kidney function is unlikely to constitute a primary risk factor for *E. coli* infection. Decision tree feature selection identified blood WBC and urine WBC as important variables (Figure 3B), and LASSO regression further revealed significant differences in urine WBC, clinical stages, urinary nitrite, blood RBC, AST, and SCC levels (Figure 3C). Based on the above results, these variables,

Table 2. The distribution characteristics of postoperative pathogenic bacterial infections in cervical cancer patients.

Pathogens	Number of Strains (n)								Total (n = 151)	Proportion (%)
	sputum	blood	urine	purulent secretion	vaginal secretion	drainage fluid	others	stool		
Gram-negative bacilli										
E. coli	1	5	61	1	1	4	3	0	76	50.33
Proteus mirabilis	0	0	10	0	1	1	1	0	13	8.61
Klebsiella pneumoniae	1	0	7	0	1	0	0	0	9	5.96
Acinetobacter baumannii	2	0	3	0	0	0	0	0	5	3.31
Pseudomonas aeruginosa	2	2	0	0	0	0	0	0	4	2.65
Citrobacter spp	0	0	0	1	0	1	0	0	2	1.32
Serratia spp	0	1	1	0	0	0	0	0	2	1.32
Klebsiella oxytoca	0	0	2	0	0	0	0	0	2	1.32
Other	1	0	2	0	0	0	0	0	3	1.99
Total	7	8	86	2	3	6	4	0	116	76.82
Gram-positive cocci										
Enterococcus faecalis	0	0	4	2	0	0	0	0	6	3.97
Staphylococcus aureus	0	1	3	0	0	1	0	0	5	3.31
Staphylococcus hominis	0	5	0	0	0	0	0	0	5	3.31
Enterococcus faecium	0	0	3	0	0	0	0	0	3	1.99
Gardnerella vaginalis	0	0	2	0	1	0	0	0	3	1.99
Streptococcus agalactiae	0	1	1	0	0	1	0	0	3	1.99
Staphylococcus epidermidis	0	1	0	0	0	0	0	0	1	0.66
Other	0	1	2	0	0	0	0	0	3	1.99
Total	0	9	15	2	1	2	0	0	29	19.21
Fungi										
Candida glabrata	1	0	1	0	0	0	0	1	3	1.99
Candida tropicalis	0	0	1	1	0	0	0	0	2	1.32
Candida albicans	1	0	0	0	0	0	0	0	1	0.66
Total	2	0	2	1	0	0	0	1	6	3.97
Total specimen count (n)	9	15	105	5	4	8	4	1	151	
Specimen proportion (%)	5.96	9.93	69.54	3.31	2.65	5.30	2.65	0.66		100

Table 3. Drug resistance in major gram-negative bacilli to commonly antimicrobial agents.

Antimicrobial agents	*E. coli* (n = 76)		*Proteus mirabilis* (n = 13)		*Klebsiella pneumoniae* (n = 9)	
	Number (n)	Resistance rate (%)	Number (n)	Resistance rate (%)	Number (n)	Resistance rate (%)
Ceftazidime	25	32.89	1	7.69	1	11.11
Piperacillin/ tazobactam	4	5.26	0	0.00	0	0.00
Tobramycin	14	18.42	2	15.38	0	0.00
Aztreonam	37	48.68	0	0.00	1	11.11
Ceftriaxone	53	69.74	6	46.15	4	44.44
Cefepime	14	18.42	0	0.00	0	0.00
Gentamicin	31	40.79	2	15.38	2	22.22
Imipenem	1	1.32	0	0.00	0	0.00
Ertapenem	1	1.32	0	0.00	0	0.00
Amikacin	7	9.21	0	0.00	0	0.00
Ciprofloxacin	50	65.79	7	53.85	4	44.44
Levofloxacin	42	55.26	3	23.08	2	22.22
Trimethoprim/ sulfamethoxazole	45	59.21	10	76.92	4	44.44
Nitrofurantoin	2	2.63	12	92.31	2	22.22

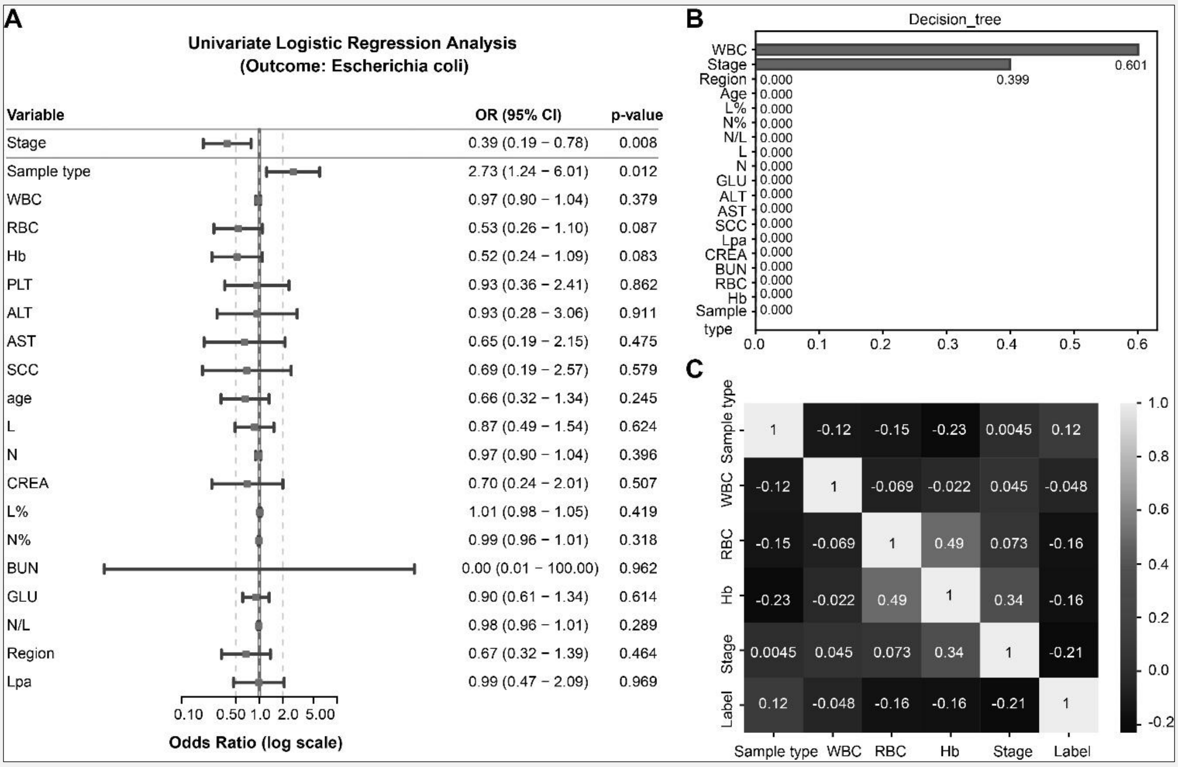
Table 4. Drug resistance in major gram-positive cocci to commonly antimicrobial agents.

Antimicrobial agents	*Enterococcus faecalis* (n = 6)		*Staphylococcus aureus* (n = 5)		*Staphylococcus hominis* (n = 5)	
	Number (n)	Resistance rate (%)	Number (n)	Resistance rate (%)	Number (n)	Resistance rate (%)
Penicillin	0	0.00	5	100.00	5	100.00
Ampicillin	0	0.00	NA	NA	NA	NA
Oxacillin	NA	NA	2	40.00	3	60.00
Vancomycin	0	0.00	0	0.00	0	0.00
Erythromycin	4	66.67	2	40.00	4	80.00
Tetracycline	5	83.33	1	20.00	0	0.00
Ciprofloxacin	0	0.00	1	20.00	1	20.00
Levofloxacin	1	16.67	1	20.00	1	20.00
Nitrofurantoin	1	16.67	0	0.00	0	0.00
Quinupristin/ dalfopristin	5	83.33	0	0.00	0	0.00
Linezolid	0	0.00	0	0.00	0	0.00
Tigecycline	0	0.00	0	0.00	0	0.00

NA indicates that antibiotic susceptibility testing was not performed for the specified pathogen against the relevant antibiotic.

including clinical stage, RBC, WBC, U_WBC1, Hb and U_nitrite, were selected for subsequent modeling.

Among the machine learning algorithms evaluated, the logistic and svm_cross models exhibited robust performance and stability. The logistic model achieved AUCs of 0.87 (training) and 0.90 (test), with corresponding sensitivities of 0.85 and 0.83, specificities of 0.79 and 0.92, positive predictive values of 0.86 and 0.94, and



Variable	OR (95% CI)	p-value
Stage	0.39 (0.19 – 0.78)	0.008
Sample type	2.73 (1.24 – 6.01)	0.012
WBC	0.97 (0.90 – 1.04)	0.379
RBC	0.53 (0.26 – 1.10)	0.087
Hb	0.52 (0.24 – 1.09)	0.083
PLT	0.93 (0.36 – 2.41)	0.862
ALT	0.93 (0.28 – 3.06)	0.911
AST	0.65 (0.19 – 2.15)	0.475
SCC	0.69 (0.19 – 2.57)	0.579
age	0.66 (0.32 – 1.34)	0.245
L	0.87 (0.49 – 1.54)	0.624
N	0.97 (0.90 – 1.04)	0.396
CREA	0.70 (0.24 – 2.01)	0.507
L%	1.01 (0.98 – 1.05)	0.419
N%	0.99 (0.96 – 1.01)	0.318
BUN	0.00 (0.01 – 100.00)	0.962
GLU	0.90 (0.61 – 1.34)	0.614
N/L	0.98 (0.96 – 1.01)	0.289
Region	0.67 (0.32 – 1.39)	0.464
Lpa	0.99 (0.47 – 2.09)	0.969

	Sample type	WBC	RBC	Hb	Stage	Label
Sample type	1	-0.12	-0.15	-0.23	0.0045	0.12
WBC	-0.12	1	-0.069	-0.022	0.045	-0.048
RBC	-0.15	-0.069	1	0.49	0.073	-0.16
Hb	-0.23	-0.022	0.49	1	0.34	-0.16
Stage	0.0045	0.045	0.073	0.34	1	-0.21
Label	0.12	-0.048	-0.16	-0.16	-0.21	1

Figure 1. Associations between *E. coli* infections and clinicopathological characteristics in all sample types.

A Univariate logistic regression evaluating the associations between *E. coli* infection and clinical variables. B Feature importance for predicting *E. coli* infection derived from decision tree analysis. C Correlation heatmap of predictor variables based on Pearson's coefficients.
L lymphocyte count, N neutrophil count, L% lymphocyte percentage, N% neutrophil percentage, N/L neutrophil-to-lymphocyte ratio.

negative predictive values of 0.78 and 0.80, respectively. The svm_cross model yielded AUCs of 0.81 (training) and 0.89 (test), with sensitivities of 0.85 and 0.78, specificities of 0.89 and 0.77, positive predictive values of 0.92 and 0.82, and negative predictive values of 0.81 and 0.71, respectively (Figure 4A and 4B). Both models maintained strong discriminatory power across the training and test sets, indicating good generalizability. No significant multicollinearity was detected (Figure 4C). A nomogram was developed to facilitate clinical risk estimation for *E. coli* infection (Figure 4D).

DISCUSSION

This retrospective study characterized the clinical and microbiological profiles of nosocomial infections in cervical cancer patients, with a specific focus on epidemiological patterns and antimicrobial resistance. Our findings indicate that infections are caused primarily by gram-negative bacteria of urinary tract origin, with *E. coli* being the most predominant pathogen. Notably, *E. coli* exhibits high resistance rates to multiple first-line antibiotics. Using machine learning approaches, we developed a prediction model for *E. coli* infection on the basis of mid-stream urine samples, which demonstrated strong discriminatory capacity (AUC > 0.8) and exhibits promising potential for clinical translation. This work provides both a theoretical basis and a practical tool for detection and precision management of nosocomial infections in culture-positive population.

Consistent with previous studies in solid tumor cohorts [10,11], gram-negative bacilli accounted for the majority of pathogens, with *E. coli* represented more than 50% of all isolates and was the most frequently recovered pathogen from mid-stream urine samples. The notion that the urinary tract is the predominant site of nosocomial infection in cervical cancer patients, is likely because of surgical intervention, indwelling catheter use, and impaired bladder function [3]. Antimicrobial susceptibility testing revealed resistance rates exceeding 55% to commonly prescribed agents including ceftriax-

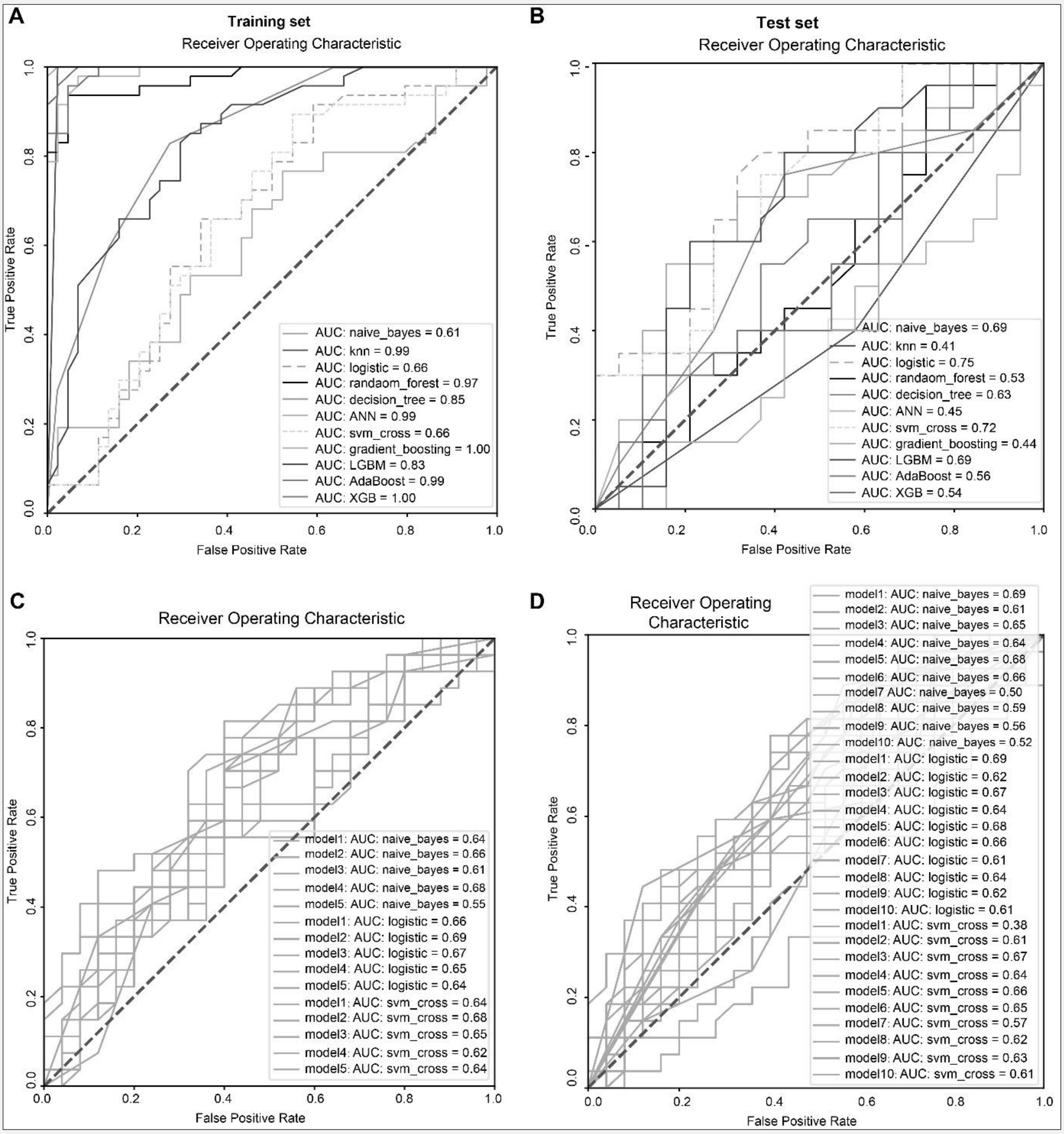


Figure 2. Development and validation of a machine learning model predicting *E. coli* infection using all sample types.

A, B ROC curves of the prediction model in the training and test sets. The predictors included the clinical stage, RBC, WBC, hemoglobin, and sample type. C ROC analysis of the naive_bayes, logistic, and svm_cross models constructed from all possible combinations of four out of the five predictors. D ROC analysis of the naive_bayes, logistic, and svm_cross models built from all combinations of the three predictors.

one and ciprofloxacin, underscoring the need for caution in empirical antibiotic selection. Conversely, the efficacy of carbapenems, β-lactam/β-lactamase inhibitor combinations, and amikacin was high, which is consistent with current national and international surveillance data [12]. These findings highlight the importance of

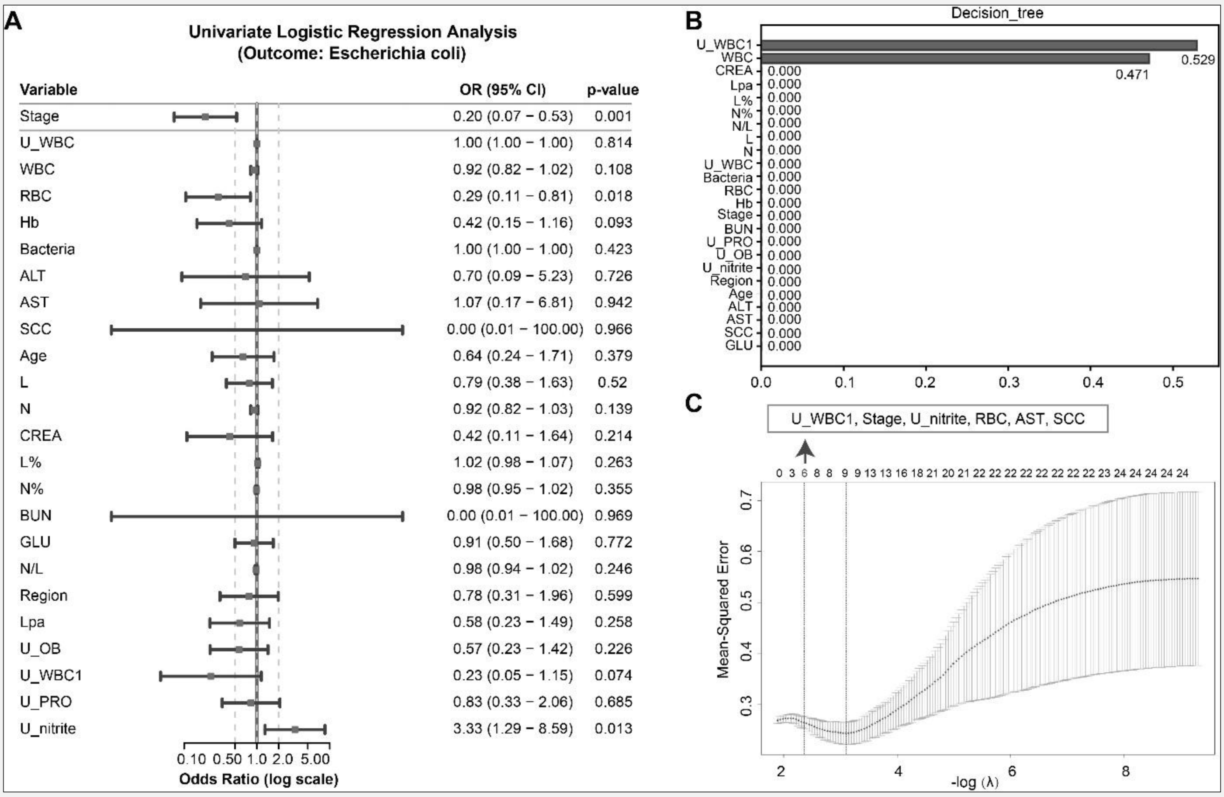


Figure 3. Relationship between *E. coli* infection and patient characteristics in mid-stream urine samples.

A Univariate logistic regression analysis of clinical factors associated with *E. coli* infection in mid-stream urine samples. B Decision tree-based feature importance ranking for *E. coli* infection prediction. C Feature selection via LASSO regression.
L lymphocyte count, N neutrophil count, L% lymphocyte percentage, N% neutrophil percentage, N/L neutrophil-to-lymphocyte ratio.
U_OB urine occult blood, U_PRO urine protein, U_WBC urine WBC count, U_WBC1 urine qualitative urine WBC, U_nitrite urine nitrite.

susceptibility-guided therapy to minimize broad-spectrum antibiotic misuse and curb the emergence of resistant strains.

Univariate logistic regression analysis revealed several factors significantly associated with *E. coli* infection, including earlier clinical stage, absence of anemia, mid-stream urine specimen type, and positive urinary nitrite status. In contrast, no significant correlations were observed with comorbid hepatic or renal dysfunction or with Lpa levels. Subsequent application of decision tree and LASSO regression methods enabled the selection of robust, non-redundant predictors, culminating in the identification of core features associated with *E. coli* infection. A notable observation was the reduced likelihood of *E. coli* infection in patients with advanced-stage anemia. We propose that this may reflect differences in treatment modalities across disease stages. Patients with advanced cancer typically undergo more radiotherapy and chemotherapy and fewer surgical procedures, thereby reducing exposure to invasive urinary interventions such as catheterization and consequent introduction of enteric pathogens such as *E. coli* [10]. Moreover, anemia in this subgroup is often related to chronic disease or myelosuppressive therapy, and these patients may benefit from more intensive clinical monitoring and empirical antibiotic coverage, potentially preventing some infections [13-15]. The overall immunocompromised state of advanced-stage patients may also predispose them to opportunistic or polymicrobial infections rather than monomicrobial *E. coli* events [16,17], possibly explaining the inverse correlation observed in our model.

For predictive modeling, we constructed *E. coli* infection models using both all sample types and only mid-stream urine samples. While algorithms such as logistic regression and support vector machines with cross-validation showed stable performance in the full sample model, their discriminative ability was limited (AUC < 0.75), likely owing to sample heterogeneity and unmea-

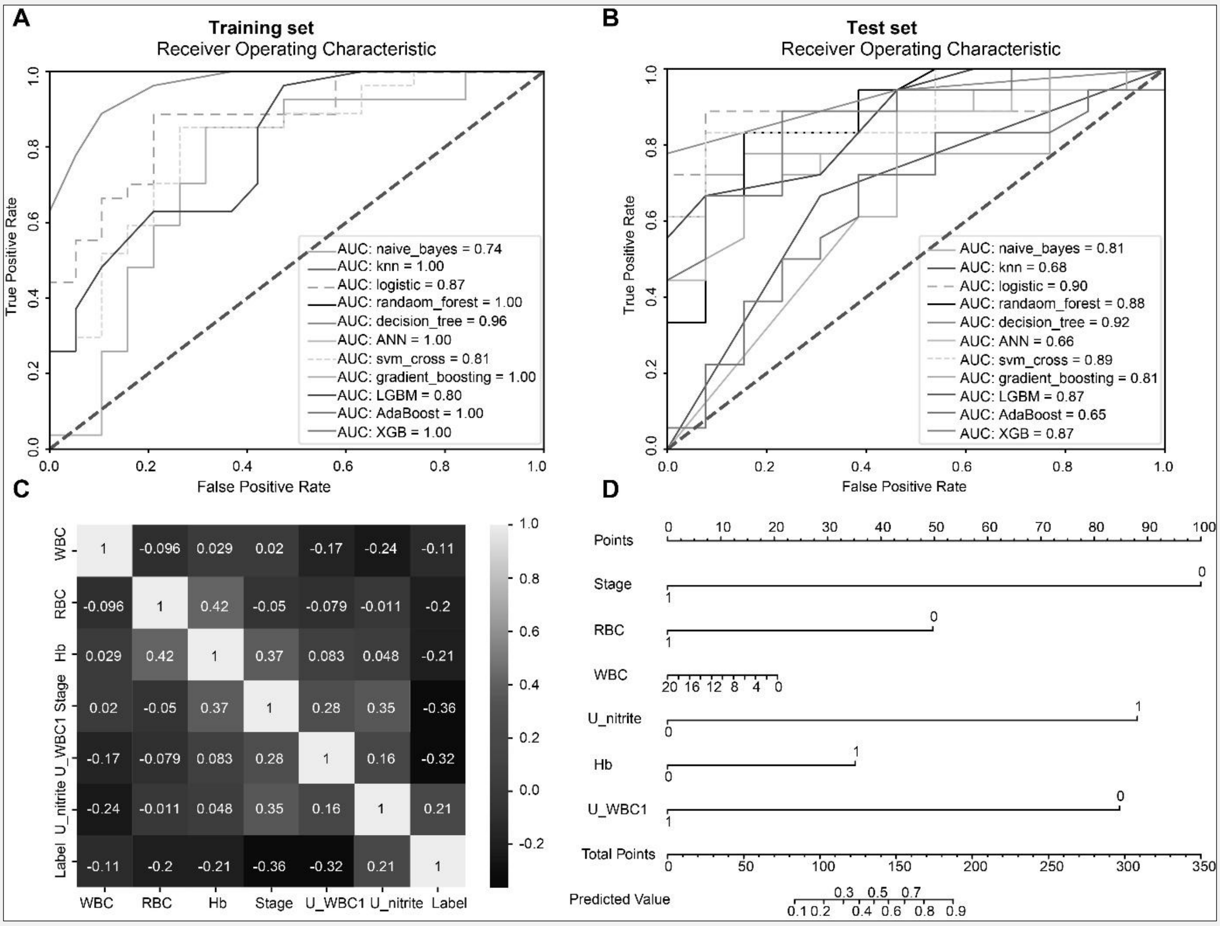


Figure 4. Performance evaluation of the machine learning prediction model for *E. coli* infection using mid-stream urine samples.

A, B ROC curve analysis demonstrating model performance in the training and test sets. The predictors included clinical stage, RBC, WBC, U_WBC1, Hb and U_nitrite. C Correlation matrix of predictor variables based on Pearson's analysis. D Clinically applicable nomogram for individualized prediction of *E. coli* infection risk.

sured confounders. Restricting the analysis to mid-stream urine samples substantially improved model performance, with the logistic and svm-cross models achieving AUCs of 0.90 and 0.89, respectively, in the test set. This underscores the value of sample homogeneity in enhancing predictive accuracy. The nomogram developed herein further augments the clinical utility and interpretability of the findings [18,19]. Recently, machine learning has gained traction in the prediction of pathogenic infections [20,21]. Prior research has employed diverse data types, from clinical and laboratory features to multiomics genomics data, to construct models for predicting infections, profiling antibiotic resistance, and screening novel therapeutics [22]. For example, Sassi et al. combined genome-wide association studies (GWASs), machine learning, and transcriptomics to predict *Staphylococcus aureus* infections [23]. Similarly, Ardila et al. leveraged whole-genome sequencing with machine learning to predict antimicrobial resistance in critical pathogens [24], and Lane et al. compared machine learning models for drug discovery against *Mycobacterium tuberculosis* [25]. Future work should prioritize the integration of multisource and multimodal data to enhance the predictive accuracy for bacterial infections and extend the applications of machine learning to elucidate resistance mechanisms and assess virulence.

Several limitations should be acknowledged. First, as a single-center retrospective study with a limited sample size, the potential for selection bias cannot be excluded.

Second, enhancing the model's accuracy and credibility requires the incorporation of diverse clinical features and parameters, such as inflammatory cytokines or microbial genomic features. Patient selection must be guided by stricter, evidence-based criteria to accurately discriminate true infections. Third, the adoption of principled approaches for handling missing data is needed to strengthen future models by minimizing bias while maintaining data completeness. Furthermore, our models face potential limitations from temporal bias and overfitting. Future studies should involve multicenter collaborations to increase the sample size, incorporate prospective validation cohorts, refine risk stratification, and integrate metagenomic sequencing and host immune parameters to develop more accurate and generalizable predictive tools.

In conclusion, this study delineates the pathogen distribution and resistance patterns of nosocomial infections in cervical cancer patients and presents a machine learning-based prediction model for *E. coli* infection using readily available clinical variables. The model exhibited excellent performance, particularly when applied to mid-stream urine samples, with strong calibration and discrimination. These findings suggest that the prediction model represents a promising tool for supporting risk stratification, guiding targeted interventions, and optimizing antibiotic therapy among culture-positive patients, with the ultimate goal of improving patient outcomes.

Acknowledgment:
We sincerely thank the MyLab + iResearch consulting team of Roche for supporting the data analysis in our study. We deeply appreciate their generosity in sharing their work and acknowledge their significant contributions.

Ethical Approval:
This study was conducted in accordance with the principles of the Declaration of Helsinki and was approved by the by the Ethics Committee of the First Affiliated Hospital of USTC (2025-RE-395). The requirement for informed consent was waived due to the retrospective nature of the study, which involved no more than minimal risk to the participants. All patient data were anonymized and deidentified prior to analysis.

Source of Support:
This work was supported by the National Natural Science Foundation of China (Grant Number: 82404092), the Natural Science Foundation of Anhui Province (Grant Number: 2408085MH203), and the Foundation of the West Branch of the First Affiliated Hospital of the University of Science and Technology of China (Grant Number: 2024YJQN007).

Declaration of Generative AI in Scientific Writing:
The authors employed the AI tool DeepSeek solely for language polishing and grammar verification to enhance readability. All the scientific content, including research design, data interpretation, and intellectual conclusions, was generated and finalized exclusively by the authors.

Declaration of Interest:
The authors declare that they have no potential conflicts of interest.

References:

1. Xu M, Cao C, Wu P, Huang X, Ma D. Advances in cervical cancer: current insights and future directions. Cancer Commun (Lond) 2025;45:77-109. (PMID: 39611440)

2. Tewari KS. Cervical Cancer. N Engl J Med 2025;392:56-71. (PMID: 39752299)

3. Jin C, Bu H, Xiang J, Jin C. Clinicopathologic and Etiologic Characteristics of Urinary Tract Infections in Patients with Cervical Cancer Undergoing Radical Operation with Indwelling Ureteral Stents. Surg Infect (Larchmt) 2025;26:453-60. (PMID: 40106237)

4. Hou Y, Pan J, Kuerban G. [Pathogenic bacterium and drug resistance in cervical cancer patients complicated with reproductive tract infection]. Zhong Nan Da Xue Xue Bao Yi Xue Ban 2016; 41:721-8. (PMID: 27592578)

5. Cui J, Zhang Y, Li X, et al. Antimicrobial resistance profiles and genome characteristics of Klebsiella isolated from the faeces of neonates in the neonatal intensive care unit. J Med Microbiol 2024;73(8):001862. (PMID: 39150452)

6. Lengert AVH, Tassinari TA, Lourenço ATO, et al. Development and evaluation of high-resolution melting assays for direct and simultaneous pathogen identification in bloodstream infections in pediatric oncology patients. Diagn Microbiol Infect Dis 2024; 110:116426. (PMID: 39163789)

7. Sarihan S, Ercan I, Saran A, Cetintas SK, Akalin H, Engin K. Evaluation of infections in non-small cell lung cancer patients treated with radiotherapy. Cancer Detect Prev 2005;29:181-8. (PMID: 15829379)

8. Luo Y, Ding W, Yang X, et al. Construction and validation of a predictive model for meningoencephalitis in pediatric scrub typhus based on machine learning algorithms. Emerg Microbes Infect 2025;14:2469651. (PMID: 39964062)

9. Villani Júnior A, Freire MP, Lazar Neto F, et al. Prediction of bacterial and fungal bloodstream infections using machine learning in patients undergoing chemotherapy. Eur J Cancer 2025;223: 115516. (PMID: 40382858)

10. Zhou M, Li H, Geng X, Dai H, Li Z. Risk Factors of Catheter-Associated Urinary Tract Infections Following Radical Hysterectomy for Cervical Cancer: A Propensity Score Matching-Based Study. Int J Womens Health 2024;16:2297-309. (PMID: 39737419)

11. Mao Y, Xu Q, Zhang J, Chou S, Shen M, Chen M. Aetiology and Prognostic Significance of Postoperative Urinary Tract Infections in Patients with Cervical Cancer. Arch Esp Urol 2024;77:1070-7. (PMID: 39632530)

12. Mubangizi L, Namusoke F, Mutyaba T. Aerobic cervical bacteriology and antibiotic sensitivity patterns in patients with advanced cervical cancer before and after radiotherapy at a national referral hospital in Uganda. Int J Gynaecol Obstet 2014;126:37-40. (PMID: 24786141)

13. Byun JM, Jeong DH. Antibiotic prophylaxis for gynecologic cancer surgery. Taiwan J Obstet Gynecol 2020;59:514-9. (PMID: 32653122)

14. Kennedy K, Gaertner-Otto J, Lim E. Reduction in deep organ-space infection in gynecologic oncology surgery with use of oral antibiotic bowel preparation: a retrospective cohort analysis. J Osteopath Med 2025;125:269-76. (PMID: 39376031)

15. De Pastena M, Paiella S, Lionetto G, et al. An Antimicrobial Stewardship Program in Pancreatic Surgery Reduces the Infectious Risk of Colonized Bile, Reducing the Predictive Value of the Intraoperative Bile Culture - A Before-after Study on 1638 Pancreatoduodenectomies. Ann Surg 2025 Nov 1;282(5):725-33. (PMID: 40747933)

16. Meng F, Zhu C, Zhu C, et al. Epidemiology and pathogen characteristics of infections following solid organ transplantation. J Appl Microbiol 2024;135(12):lxae292. (PMID: 39567858)

17. Garg P, Singh N, Liu AJ, et al. Estimating the prevalence of key healthcare-associated and opportunistic infections in Australian transplant and cancer populations: protocol for the PROSPER point prevalence study. BMJ Open 2025;15:e100798. (PMID: 40750274)

18. Du W, Ji W, Luo T, et al. Development of a Prognostic Nomogram for Nonneutropenic Invasive Pulmonary Aspergillosis Based on Machine Learning. J Inflamm Res 2024;17:9823-35. (PMID: 39618929)

19. Sun T, Liu J, Yuan H, et al. Construction of a risk prediction model for lung infection after chemotherapy in lung cancer patients based on the machine learning algorithm. Front Oncol 2024;14:1403392. (PMID: 39184040)

20. Zhao Q, Liu MY, Gao KX, et al. Predicting 90-day risk of urinary tract infections following urostomy in bladder cancer patients using machine learning and explainability. Sci Rep 2025;15:6807. (PMID: 40000794)

21. Gotti C, Roux-Dalvai F, Bérubé È, et al. LC-SRM Combined With Machine Learning Enables Fast Identification and Quantification of Bacterial Pathogens in Urinary Tract Infections. Mol Cell Proteomics 2024;23:100832. (PMID: 39178943)

22. Abhadionmhen AO, Asogwa CN, Ezema ME, et al. Machine Learning Approaches for Microorganism Identification, Virulence Assessment, and Antimicrobial Susceptibility Evaluation Using DNA Sequencing Methods: A Systematic Review. Mol Biotechnol Epub 2024 Nov 9. (PMID: 39520638)

23. Sassi M, Bronsard J, Pascreau G, et al. Forecasting Staphylococcus aureus Infections Using Genome-Wide Association Studies, Machine Learning, and Transcriptomic Approaches. mSystems 2022;7:e0037822. (PMID: 35862809)

24. Ardila CM, Yadalam PK, González-Arroyave D. Integrating whole genome sequencing and machine learning for predicting antimicrobial resistance in critical pathogens: a systematic review of antimicrobial susceptibility tests. PeerJ 2024;12:e18213. (PMID: 39399439)

25. Lane T, Russo DP, Zorn KM, et al. Comparing and Validating Machine Learning Models for Mycobacterium tuberculosis Drug Discovery. Mol Pharm 2018;15:4346-60. (PMID: 29672063)