Adaptation and External Validation of Pathogenic Urine Culture Prediction in Primary Care Using Machine Learning

Gurpreet Dhanda, MD
Mirna Asham, MD
Denton Shanks, DO, MPH
Nicole O'Malley, MD
Joel Hake, MD
Megha Teeka Satyan, MD
Nicole T. Yedlinsky, MD
Daniel J. Parente, MD, PhD

Department of Family Medicine and Community Health, University of Kansas Medical Center, Kansas City, Kansas

MORE ONLINE
www.annfammed.org

Conflicts of interest: authors report none.

CORRESPONDING AUTHOR

Daniel J. Parente
Department of Family Medicine
and Community Health
University of Kansas Medical Center
3901 Rainbow Blvd, MS 4010
Kansas City, KS 66160
dparente@kumc.edu

ABSTRACT

BACKGROUND Urinary tract infection (UTI) symptoms are common in primary care, but antibiotics are appropriate only when an infection is present. Urine culture is the reference standard test for infection, but results take >1 day. A machine learning predictor of urine cultures showed high accuracy for an emergency department (ED) population but required urine microscopy features that are not routinely available in primary care (the NeedMicro classifier).

METHODS We redesigned a classifier (NoMicro) that does not depend on urine microscopy and retrospectively validated it internally (ED data set) and externally (on a newly curated primary care [PC] data set) using a multicenter approach including 80,387 (ED) and 472 (PC) adults. We constructed machine learning models using extreme gradient boosting (XGBoost), artificial neural networks, and random forests (RFs). The primary outcome was pathogenic urine culture growing ≥100,000 colony forming units. Predictor variables included age; gender; dipstick urinalysis nitrites, leukocytes, clarity, glucose, protein, and blood; dysuria; abdominal pain; and history of UTI.

RESULTS Removal of microscopy features did not severely compromise performance under internal validation: NoMicro/XGBoost receiver operating characteristic area under the curve (ROC-AUC) 0.86 (95% CI, 0.86-0.87) vs NeedMicro 0.88 (95% CI, 0.87-0.88). Excellent performance in external (PC) validation was also observed: NoMicro/RF ROC-AUC 0.85 (95% CI, 0.81-0.89). Retrospective simulation suggested that NoMicro/RF can be used to safely withhold antibiotics for low-risk patients, thereby avoiding antibiotic overuse.

CONCLUSIONS The NoMicro classifier appears appropriate for PC. Prospective trials to adjudicate the balance of benefits and harms of using the NoMicro classifier are appropriate.

Ann Fam Med 2023;21:11-18. https://doi.org/10.1370/afm.2902

INTRODUCTION

Urinary tract infections (UTIs) are the most common type of infections managed in the outpatient setting, accounting for 1% to 3% of all consultations, 15% of all community prescriptions for antibiotics, and $1.6 billion in annual health care costs.[1,2] Both men and women can be affected. Women have a 50% to 60% lifetime risk of UTI. Women >65 years of age are affected twice as often as the general female population.[1] Surveys indicate that primary care (PC) clinicians are concerned about antibiotic resistance, but many view this as a public health issue in general rather than as a factor in prescribing decisions for individual patients.[3]

Urinary tract infection is usually diagnosed by combining history and physical examination with urine dipstick testing (including nitrite and leukocyte esterase).[4,5] Microscopic evaluation of the urine to identify the presence of, for example, bacteria, leukocytes, and squamous epithelial cells is sometimes performed but is not always immediately available in the outpatient setting.

Urine culture is the reference standard for UTI diagnosis; however, urine cultures often take ≥24 hours for results whereas antibiotic treatment decisions are often made in minutes—during an office visit—at the point of care. Accurate prediction of urine cultures could enable prompt treatment of patients with UTI while avoiding antibiotic overuse for those without UTI.

Several approaches to more accurate diagnosis and treatment of UTI have been developed.[6-11] Little et al[7] developed a dipstick rule—based on the presence of nitrite or both leukocytes and blood—with a sensitivity of 77%, a specificity of

70%, and a negative predictive value (NPV) of 65%. Likewise, McIsaac et al[8] developed a 3-variable decision aid (dysuria, leukocytes, nitrites) with a sensitivity of 80.3% and a specificity of 53.7%.

More recently, machine learning algorithms have been devised to predict the outcome of urine cultures.[6,11] Heckerling et al[6] used artificial neural networks (ANNs)[12] to produce 5-variable predictors of urine culture results for a small data set (212 women). In a more recent study—using a much larger data set of >80,000 emergency department (ED) encounters—Taylor et al[11] were able to predict the pathogenicity of a urine culture with high discriminative performance (reported receiver operating characteristic area under the curve [ROC-AUC] of 0.904 for the full [but impractical] 212-variable model and 0.877 for the reduced [but practical] 10-variable model). Their predictor leveraged a new machine learning approach based on extreme gradient boosting (XGBoost).[11,13] The Taylor approach modeled the presence of a urine culture as a function of the following 10 variables: 2 demographic features (age, gender/sex), 3 urine dipstick features (nitrites, leukocyte esterase, blood), 2 history features (presence of dysuria, history of UTI), and 3 urine microscopy features (bacteria, epithelial cells, leukocytes).[11]

Unfortunately, in many ambulatory PC settings (eg, family medicine offices or urgent care facilities), urine microscopy is not immediately available, and treatment decisions are often made without this information. Urine microscopy provides valuable information for the evaluation of the pathogenicity of a urine culture. The presence of white blood cells and bacteria argues in favor of infection. Detection of squamous epithelial cells is a quality-control marker that suggests contamination with commensal urogenital flora. A risk therefore exists that removal of microscopic features from the prediction model might severely compromise performance.

We investigated whether the Taylor[11] model could be adapted to remove the dependence on urine microscopy without compromising predictive accuracy and whether a model built on ED encounters could be generalized to PC patients at a different medical center. To that end, we developed a new model (NoMicro) that does not depend on urine microscopy variables. We trained and internally validated this new model on the original ED data set and then externally validated it on a new data set of 472 outpatient encounters in a family medicine office at a different institution.

Table 1. Data Source Demographic Characteristics

Characteristic	Primary Care	Emergency Department		
		Total	Training	Validation
No.	472	80,387	64,310	16,077
Urine culture pathogenicity, No. (%)				
Pathogenic	128 (27.1)	18,284 (22.7)	14,718 (22.9)	3,566 (22.2)
Nonpathogenic	344 (72.9)	62,103 (77.3)	49,592 (77.1)	12,511 (77.8)
Age, y, No. (%)				
18-25	51 (10.8)	10,052 (12.5)	8,077 (12.6)	1,975 (12.3)
26-35	87 (18.4)	11,891 (14.8)	9,455 (14.7)	2,436 (15.2)
36-45	85 (18.0)	9,450 (11.8)	7,525 (11.7)	1,925 (12.0)
46-55	59 (12.5)	12,255 (15.2)	9,825 (15.3)	2,430 (15.1)
56-65	90 (19.1)	10,327 (12.8)	8,230 (12.8)	2,097 (13.0)
66-75	67 (14.2)	9,214 (11.5)	7,380 (11.5)	1,834 (11.4)
>75	33 (7.0)	17,198 (21.4)	13,818 (21.5)	3,380 (21.0)
Gender, No. (%)				
Male	64 (13.6)	24,584 (31.0)	19,648 (31.0)	4,936 (31.1)
Female	408 (86.4)	54,725 (69.0)	43,803 (69.0)	10,922 (68.9)
Not reported	NA	1,078	859	219
Race, No. (%)				
Asian	23 (4.9)	860 (1.1)	688 (1.1)	172 (1.1)
Black	160 (34.0)	17,003 (21.9)	13,541 (21.8)	3,462 (22.3)
White	211 (44.8)	43,156 (55.5)	34,596 (55.6)	8,560 (55.1)
Other/multiple	77 (16.3)	16,735 (21.5)	13,402 (21.5)	3,333 (21.5)
Not reported	1	2,633	2,083	550
Ethnicity, No. (%)				
Hispanic, Latine, Spanish origin	58 (12.3)	17,064 (21.6)	13,634 (21.6)	3,430 (21.7)
Not Hispanic, Latine, Spanish origin	412 (87.7)	61,826 (78.4)	49,474 (78.4)	12,352 (78.3)
Not reported	2	1,497	1,202	295

Note: Percentages reflect the proportion of reported values (ie, excluding not reported).

METHODS

Data Sources

We used the following 2 data sources: a sample of >80,000 patients seen in an ED and previously described by Taylor et al[11] (the ED data set), and a sample of 472 patients seen at the outpatient family medicine department at the University of Kansas Medical Center (the PC data set). Data extraction and quality assurance for the PC data set is detailed in Supplemental Appendix 1. The ED data set was further divided into training (80%, n = 64,310) and internal validation (20%, n = 16,077) data sets. The PC data set was used exclusively for external validation (ie, not for training). Characteristics of these data sets are summarized in Table 1.

Model Specification and Training

We trained urine culture predictive models using R v.3.6.1 (The R Foundation)

using software and methods described by Taylor et al.[11] The microscopy-required (NeedMicro) model was the Taylor et al[11] model, specified as follows:

NeedMicro model, pathogenic culture = Age + Gender + (Nitrite * Leukocytes) + Blood + Dysuria + History of UTIs + (Microscopic bacteria * Microscopic epithelial cells) + Microscopic white blood cells

The microscopy-independent (NoMicro) model, using only data likely to be available during a PC office visit, was specified as follows:

NoMicro model, pathogenic culture = Age + Gender + Nitrite + Leukocytes + Blood + Clarity + Glucose + Protein + Dysuria + Abdominal pain + History of UTIs

To compensate for the loss of microscopy information, we added 4 features to the NoMicro model that might be (positively or negatively) associated with UTI, compared with relevant differential diagnoses, including 3 dipstick urinalysis features (clarity, glucose, protein) and 1 history feature (abdominal pain). These features can all be readily measured during a PC office visit.

We trained the NoMicro models using XGBoost,[13] random forests (RFs),[14,15] and ANNs.[12] The NeedMicro model has been shown to perform best using XGBoost, and we trained this model using XGBoost as described.[11]

Model Validation

We internally validated trained models using the emergency department 20% holdout validation set and externally validated on the primary care data set. First, we determined the overall discriminative performance (ROC-AUC) and scaled Brier score. Clinical use of the classifiers depends not on their overall performance but on their performance at specific cutoffs for prediction of pathogenic and nonpathogenic; above the cutoff, cultures are predicted to be pathogenic, and below the cutoff, cultures are predicted to be non pathogenic. By varying the cutoff, greater sensitivity can be achieved in exchange for less specificity (and vice versa). We therefore characterized the sensitivity, specificity, positive predictive value (PPV), NPV, positive likelihood ratio, negative likelihood ratio, and diagnostic odds ratio at the optimal cutoff (ie, the cutoff maximizing the Youden index [sensitivity + specificity − 1]) and at a 15% false-negative rate (85% sensitivity [Sen85]). The significance of the 15% false-negative rate cutoff is that it allows for understanding of how the model will perform when used to reliably infer the absence of a pathogenic culture, as might be useful in supporting a decision to defer empiric antibiotic use (thereby decreasing antibiotic overuse). Finally, we determined model calibration; for example, a culture with a pathogenicity prediction of 30% should turn out to actually be pathogenic approximately 30% of the time. Further details of model validation can be found in Supplemental Appendix 1 Methods.[16]

Retrospective Evaluation of Potential to Decrease Antibiotic Overuse

We retrospectively evaluated the clinical effect of applying the following decision rules to the NoMicro models on the PC data set:

Rule 1. For patients with a culture predicted to be nonpathogenic under the Sen85 cutoff, we simulated the effect of withholding antibiotics.

Rule 2. For patients for whom the model predicts a pathogenic culture (at the Sen85 cutoff), we left the provision of antibiotics to physician discretion.

Our approach was to identify situations in which the models suggest antibiotics might be safely deferred to decrease antibiotic overuse. To mimic the population of a prospective clinical trial as closely as possible, we included all nonpregnant adults and excluded any individuals with high-risk features (eg, those suggestive of sepsis or pyelonephritis; see Supplemental Appendix 1).

Data and Source Code Availability

The deidentified PC data set and the statistical analysis code are available at https://github.com/djparente/uti-ml.

Human Subjects Protection

The University of Kansas Medical Center Institutional Review Board approved this project.

RESULTS

Comparison Between Emergency Department and Primary Care Data Sets

Demographic features of the ED and PC data sets are shown in Table 1. The ED data set comprised 80,387 individuals, whereas the PC data set comprised 472 individuals. Cultures were slightly more likely to be nonpathogenic in the ED data set (77.3%) compared with the PC data set (72.9%). Relative to PC patients, ED patients were more commonly older (32.9% aged >65 years vs 21.2% for PC patients), male (31.0% vs 13.6%), and of Hispanic/Latine/Spanish ethnicity (21.6% vs 12.3%). Racial distributions were broadly similar, although with a greater proportion of non-White patients in the PC data set (55.2%) compared with the ED data set (44.5%). The distribution of demographic and model predictor variables, stratified by urine culture pathogenicity, are reported in Supplemental Table 1.

Internal Validation of the Redesigned Classifier to Eliminate Dependence on Microscopy Data

We compared the redesigned (NoMicro) classifier with the original (NeedMicro) classifier using the ED data set (internal validation). First, we evaluated overall performance (Table 2). The NoMicro classifier using XGBoost, RFs, and ANNs were evaluated against the best-performing NeedMicro classifier (which was trained using XGBoost[11]). For concision, we

refer to, for example, the NoMicro classifier trained using the XGBoost algorithm as NoMicro/XGBoost. The 3 NoMicro classifiers performed similarly to each other and to the NeedMicro classifier (Table 2, Table 3, and Figure 1a). The best NoMicro classifier was achieved using XGBoost (as was the case for the NeedMicro classifier previously[11]). The ROC-AUC for the NoMicro/XGBoost classifier was 0.86 (95% CI, 0.86-0.87), and the NeedMicro classifier achieved an

Table 2. Discriminative Performance (ROC-AUC), Calibration, and Brier Scores for the NoMicro and NeedMicro Predictive Models Under Internal (Emergency Department) and External (Primary Care) Validation

	ROC-AUC (95% CI[a])		Calibration Decile Linear Fit R^2 (95% CI[a])		Scaled Brier Score (95% CI[a])	
Model	**Primary Care[b]**	**Emergency Department[c]**	**Primary Care[b]**	**Emergency Department[c]**	**Primary Care[b]**	**Emergency Department[c]**
NoMicro/XGB	0.84 (0.8-0.88)	0.86 (0.86-0.87)	0.98 (0.83-0.98)	>0.99 (0.99-1.0)	0.34 (0.25-0.42)	0.34 (0.33-0.36)
NoMicro/RF	0.85 (0.81-0.89)	0.85 (0.84-0.85)	0.94 (0.77-0.97)	>0.99 (0.98-1.0)	0.37 (0.27-0.46)	0.3 (0.28-0.32)
NoMicro/ANN	0.85 (0.81-0.89)	0.86 (0.85-0.86)	0.97 (0.86-0.98)	>0.99 (0.99-1.0)	0.35 (0.26-0.43)	0.33 (0.32-0.35)
NeedMicro/XGB	NA[d]	0.88 (0.87-0.88)	NA[d]	>0.99 (0.99-1.0)	NA[d]	0.4 (0.38-0.42)

ANN = artificial neural networks; AUC = area under the curve; NA = not applicable; R^2 = coefficient of determination; RF = random forests; ROC = receiver operating characteristic; XGB = extreme gradient boosting (XGBoost).

[a] Estimate and 95% CI values across 2,000 stratified (by pathogenicity) bootstrap replicates using the percentage method.
[b] External validation on the primary care data set.
[c] Internal validation on the emergency department data set.
[d] The NeedMicro classifier cannot be validated on the primary care data set because urine microscopy data are not available for almost all records.

Table 3. Cutoff-Varying Performance Metrics: Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value, Likelihood Ratios, and Diagnostic Odds Ratio

		Performance Metric Estimate, % (95% CI[a])			
Model	**Threshold**	**Sensitivity**	**Specificity**	**PPV**	**NPV**
External validation: primary care data set					
NoMicro/XGB	Best	72.7 (64.8-80.5)	82.8 (78.8-86.9)	61.2 (55.3-67.7)	89.1 (86.2-92.0)
NoMicro/RF	Best	78.9 (71.9-85.2)	81.4 (77.6-85.5)	61.2 (56.0-67.3)	91.2 (88.4-93.8)
NoMicro/ANN	Best	78.1 (71.1-85.2)	78.2 (73.5-82.6)	57.1 (51.8-62.7)	90.6 (87.8-93.3)
NoMicro/XGB	Sen85	85.2 (78.9-90.6)	62.8 (57.6-68.0)	46.0 (42.5-50.0)	91.9 (88.9-95.0)
NoMicro/RF	Sen85	85.2 (78.9-90.6)	66.0 (60.8-70.9)	48.2 (44.1-52.6)	92.3 (89.1-95.1)
NoMicro/ANN	Sen85	85.2 (78.9-90.6)	59.6 (54.1-64.5)	44.0 (40.3-47.7)	91.5 (88.1-94.7)
Internal validation: emergency department data set					
NoMicro/XGB	Best	80.0 (78.7-81.3)	76.3 (75.6-77.1)	49.1 (48.2-50.0)	93.0 (92.6-93.5)
NoMicro/RF	Best	70.6 (69.1-72.0)	83.1 (82.4-83.8)	54.4 (53.2-55.5)	90.8 (90.4-91.3)
NoMicro/ANN	Best	78.6 (77.2-79.9)	77.3 (76.6-78.1)	49.7 (48.8-50.6)	92.7 (92.2-93.1)
NeedMicro/XGB	Best	76.1 (74.6-77.5)	83.7 (83.0-84.3)	57.1 (56.0-58.1)	92.5 (92.0-92.9)
NoMicro/XGB	Sen85	85.0 (83.8-86.1)	70.5 (69.7-71.3)	45.1 (44.3-45.8)	94.3 (93.9-94.7)
NoMicro/RF	Sen85	85.1 (83.9-86.2)	64.4 (63.6-65.3)	40.6 (39.9-41.2)	93.8 (93.3-94.3)
NoMicro/ANN	Sen85	85.0 (83.8-86.2)	69.5 (68.7-70.3)	44.3 (43.5-45.0)	94.2 (93.8-94.7)
NeedMicro/XGB	Sen85	85.0 (83.8-86.2)	73.1 (72.4-73.9)	47.4 (46.6-48.2)	94.5 (94.1-94.9)

ANN = artificial neural networks; Best = threshold maximizing the Youden index (sensitivity + specificity – 1); DOR = diagnostic odds ratio (ratio of LR+ to LR–); LR– = negative likelihood ratio; LR+ = positive likelihood ratio; NPV = negative predictive value; PPV = positive predictive value; RF = random forests; Sen85 = threshold obtained by requiring the greatest specificity such that sensitivity is >85% (ie, false negative rate is <15%); XGB = extreme gradient boosting (XGBoost).

[a] Estimate and 95% CI values across 2,000 stratified bootstrap replicates using the percentage method.

ROC-AUC of 0.88 (95% CI, 0.87-0.88; precisely in concordance with Taylor et al's[11] result).

Next, we evaluated performance measures at 2 prediction cutoffs (Table 3). At the optimal cutoff, the best NoMicro classifier (NoMicro/XGBoost) achieved superior sensitivity (80.0%; 95% CI, 78.7%-81.3%) compared with the NeedMicro classifier (76.1%; 95% CI, 74.6%-77.5%), at the cost of specificity (NoMicro/XGBoost: 76.3%; 95% CI, 75.6%-77.1% vs NeedMicro: 83.7%; 95% CI, 83.0%-84.3%). At the Sen85 cutoff, the NoMicro and NeedMicro classifiers achieved similar specificity (NoMicro/XGBoost: 70.5%; 95% CI, 69.7%-71.3% vs NeedMicro: 73.1%; 95% CI, 72.4%-73.9%). At this threshold, both models also had excellent NPV (NoMicro/XGBoost: 94.3%; 95% CI, 93.9%-94.7% vs NeedMicro: 94.5%; 95% CI, 94.1%-94.9%).

We further evaluated the calibration of the models. All models were well calibrated on the ED validation data set (Table 2, Supplemental Table 2, and Figure 1b). Linear fits of the observed vs predicted pathogenicity rate within risk deciles (Supplemental Table 2) explained most of the variability in the decile plots ($R^2 \geq 0.995$ for all fits). All models—except NoMicro/RF—also produced decile-plot fits with slopes and intercepts approximately equal to 1 and 0, respectively, as expected. The pattern of residuals around the 45° perfect-calibration line in the decile plots also showed no systematic deviation, except in the case of NoMicro/RF, which underestimated the pathogenicity of predicted-to-be-pathogenic cultures and overestimated the pathogenicity of predicted-to-be-benign cultures.

Performance Metric Estimate, % (95% CI[a])		
LR+	LR–	DOR
4.24 (3.32-5.62)	0.33 (0.24-0.43)	12.8 (8.1-21.5)
4.24 (3.42-5.53)	0.26 (0.18-0.35)	16.4 (10.2-28.4)
3.58 (2.89-4.52)	0.28 (0.19-0.37)	12.8 (8.3-21.5)
2.29 (1.99-2.69)	0.24 (0.14-0.34)	9.7 (6.1-17.9)
2.50 (2.12-2.98)	0.23 (0.14-0.33)	11.1 (6.6-20.0)
2.11 (1.82-2.45)	0.25 (0.15-0.36)	8.5 (5.1-15.5)
3.38 (3.27-3.50)	0.26 (0.25-0.28)	12.9 (11.7-14.2)
4.18 (3.99-4.38)	0.35 (0.34-0.37)	11.8 (10.8-12.9)
3.47 (3.35-3.59)	0.28 (0.26-0.3)	12.5 (11.5-13.7)
4.66 (4.47-4.87)	0.29 (0.27-0.3)	16.3 (14.9-17.8)
2.88 (2.79-2.97)	0.21 (0.2-0.23)	13.6 (12.3-15.0)
2.39 (2.33-2.46)	0.23 (0.21-0.25)	10.3 (9.4-11.4)
2.79 (2.71-2.87)	0.22 (0.2-0.23)	12.9 (11.7-14.3)
3.17 (3.07-3.27)	0.21 (0.19-0.22)	15.5 (14.1-17.1)

External Validation on the Primary Care Data Set

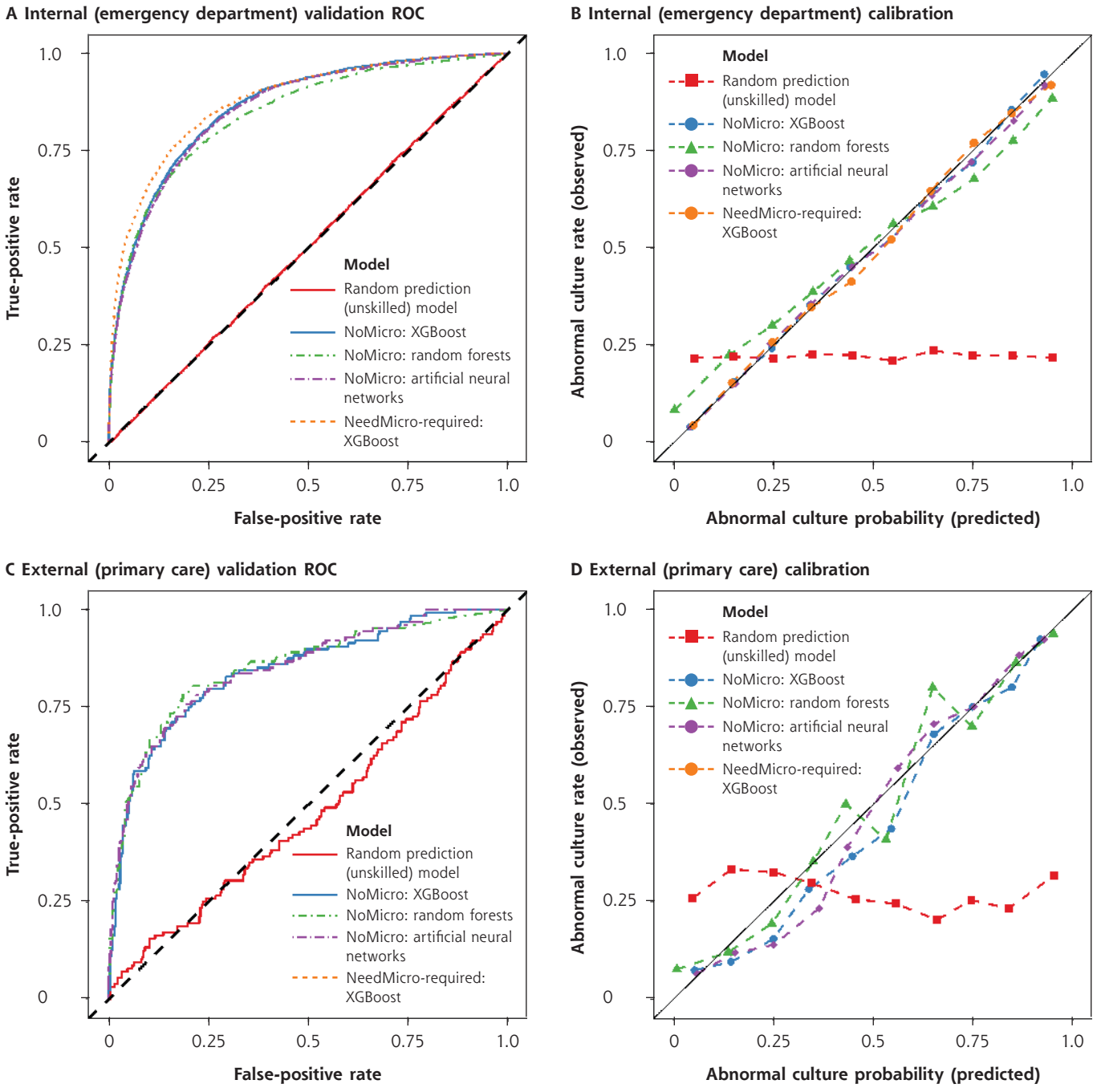
We next determined whether the NoMicro classifiers would perform adequately in our clinical setting of interest (PC rather than ED) at a different institution (external validation). The NoMicro classifiers all performed excellently on this data set (Table 2, Table 3, and Figure 1c). The best NoMicro classifier, NoMicro/RF, achieved an overall ROC-AUC of 0.85 (95% CI, 0.81-0.89). At the optimal threshold, NoMicro/RF had a sensitivity of 78.9% (95% CI, 71.9%-85.2%) and a specificity of 81.4% (95% CI, 77.6%-85.5%). At the Sen85 threshold, NoMicro/RF had a specificity of 66.0% (95% CI, 60.8%-70.9%). Importantly, the NPV at the Sen85 threshold was 92.3% (95% CI, 89.1%-95.1%).

Regarding calibration, the pattern of fluctuations above and below the ideal calibration line (a line with slope 1 and intercept 0) showed no systemic deviation, indicating that none of the NoMicro models systematically overestimated or underestimated predictions. Indeed, all models—this time including NoMicro/RF—also produced decile-plot fits with slopes and intercepts approximately equal to 1 and 0, respectively, as expected, on the PC data set. However, within risk deciles, there was greater variability, with corresponding lower—but still high—R^2 values (Table 2, R^2 values from 0.94 to 0.98, compared with the ED data set, with R^2 >0.99 for all models). Note that we could not evaluate the NeedMicro classifier on the PC data set because nearly all (~90%) PC encounters did not include microscopy data (indeed, this was the original rationale for the development of the NoMicro model).

Potential to Decrease Antibiotic Overuse

We evaluated whether the NoMicro model could be used to potentially decrease antibiotic overuse (Figure 2, Supplemental Figure 1, and Supplemental Figure 2). Of 472 included encounters, 219 had ≥1 high-risk feature, whereas 253 lacked high-risk features. Of the 253 individuals lacking these features, the NoMicro/RF model (Figure 2) predicted 119 cultures to be pathogenic and 134 to be nonpathogenic. In the predicted-pathogenic arm, the decision rule (Rule 2) recommends no change to antibiotic decision making. In the predicted-nonpathogenic arm, there were 15 instances in which the decision rule (Rule 1) would have recommended withholding antibiotics in situations in which physicians (without the benefit of the decision rule) prescribed antibiotics. Almost all decisions to prescribe antibiotics contrary to the decision rule's recommendation to withhold them would have been incorrect; 14 of the 15 instances (93.3%) had

Figure 1. Discriminative performance and calibration of models under internal and external validation.

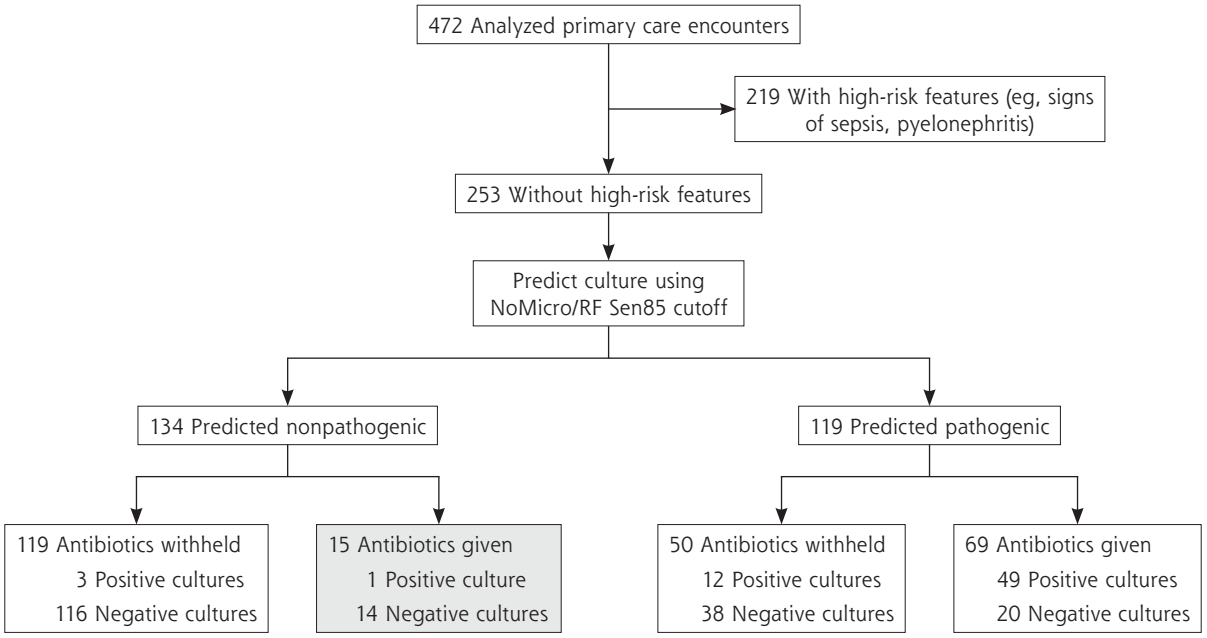


ROC = receiver operating characteristic; XGBoost = extreme gradient boosting.

Note: ROC (panels A and C) and calibration curves (panels B and D) for internal (emergency department, panels A-B) and external (primary care, panels C-D) validations. For internal validation, the NoMicro and NeedMicro models were evaluated. For external validation, only the NoMicro models were evaluated because microscopy is not routinely available in the primary care setting. Better models have ROC curves deflected away from the midline and toward the upper left corner. Well-calibrated models should lie along the diagonal line. The performance of unskilled classifiers (which return random results) were also simulated and are shown for comparison.

negative urine cultures, whereas only 1 instance had a positive culture. Similar results were obtained using the other classifiers NoMicro/XGBoost (Supplemental Figure 1) and NoMicro/ANN (Supplemental Figure 2). For NoMicro/RF, the decision rules would have increased the incidence for which antibiotics were (correctly) withheld from patients with negative urine cultures from 81.9% to 89.4% (+7.4%) while only increasing the incidence for which antibiotics were (incorrectly) withheld from patients with positive urine cultures from 23.1% to 24.6% (+1.5%).

Figure 2. Evaluation of the potential of NoMicro to decrease antibiotic overuse (random forests).



RF = random forests; Sen85 = threshold obtained by requiring the greatest specificity such that sensitivity is >85% (ie, false-negative rate is <15%).

Note: Of 472 primary care encounters, 253 lacked high-risk features for progression to serious illness and were stratified using the NoMicro/Random Forests classifier at the Sen85 cutoff (false-negative rate 15%). These predictions were correlated with physician antibiotic prescribing behavior (made without the benefit of the NoMicro/RF model). The shaded box represents cases for which the NoMicro/RF model predicts the culture to be nonpathogenic but for which physicians nevertheless prescribed antibiotics; almost all cultures in this group were negative. Antibiotic overuse might be decreased by withholding antibiotics for this group.

DISCUSSION

We investigated whether a previously successful urine culture prediction model[11] in an ED could be adapted to remove its dependence on microscopy data (the NoMicro model), thereby rendering it appropriate for environments that lack the ability to characterize urine microscopically at the point of care (eg, PC or urgent care). In internal validation (ED data set), a statistical difference between the ROC-AUC for the NoMicro/XGBoost and NeedMicro classifiers was found (DeLong test $P < .001$). However, although the large sample size allowed this statistical difference to be detected, this difference is unlikely to have major clinical implications; both the NoMicro and NeedMicro classifiers achieved high performance (ROC-AUC both >0.85). Likewise, cutoff-dependent performance measures (sensitivity, specificity, PPV, and NPV) were broadly comparable in the NoMicro and NeedMicro models. Taken together, the overall and cutoff-dependent performance under internal validation suggest that the NoMicro classifiers are viable alternatives to the NeedMicro classifier and are not severely impaired by the loss of urine microscopy features from the prediction model.

Similarly, performance statistics under external validation—in a different clinical setting (PC) at a different institution—were similar to those obtained during internal validation. This strongly suggests that removal of microscopic features does not substantially impair performance of the prediction model, the model is not significantly overfit to the peculiarities of the ED data set, and the use of the NoMicro classifiers in PC populations is reasonable and generalizable across at least some institutions and practice settings.

Specifically, our results establish the validity of the NoMicro model in a single-center ED (internal validation) and a different, but still single-center, PC environment (external validation) at an urban academic medical center. That the NoMicro model—which was trained on the ED data set—works well in a completely different clinical setting (PC) and physical location suggests that the model is generalizable. Our results nevertheless do not definitively establish this. It remains formally possible that the model might not be valid in settings with more profound differences (eg, an affluent, suburban, community urgent care). Future studies might investigate this.

At both the optimal and Sen85 cutoffs, the NPV of the NoMicro/RF model was excellent (>90%). However, the PPV was much lower (61.2%; 95% CI, 56.0%-67.3% at the optimal cutoff and 48.2%; 95% CI, 44.1%-52.6% at the Sen85 cutoff). Our model might therefore be useful for UTI in a manner that is analogous to how a D-dimer test might be used to rule out pulmonary embolism[17]; useful to withhold antibiotics (reasonably exclude infection) when negative but not useful to infer an infection when positive. (That is, antibiotics should not be started solely based on a pathogenic prediction from the NoMicro model). Our results therefore suggest that use of the proposed decision rules could decrease antibiotic

overuse without substantially withholding antibiotics in the setting of an infection.

The NoMicro model is more complicated than scoring system–type decision rules that simply add up the number of risk factors present and compare them to a prespecified cutoff. Custom software—preferably web-based, for broad availability—will need to be written and validated to allow clinicians to act on NoMicro predictions. Importantly, such software should not describe predicted-pathogenic results as high risk or potentially pathogenic, which might unintentionally cause physicians to prescribe antibiotics when they otherwise might have chosen to defer them. Development and validation of such a tool is a major future direction of this work. Among the goals of such work would be to understand under what conditions the NoMicro predictions (the NoMicro model is essentially a black-box model) are viewed as acceptable to PC physicians.

The present study has limitations. First, the PC data set was relatively small (n = 472) and described a single center. The total number of individuals prescribed antibiotics, for whom the proposed decision rules suggested antibiotics be withheld, was correspondingly small (15 individuals). Second, the ED and PC data sets had a pathogenic urine culture prevalence of ~25%. Measured ROC-AUC, sensitivity, and specificity do not change with population prevalence, but NPV does. In a practice setting with a greater prevalence of pathogenic urine cultures, the NoMicro NPV will be lower. Third, our data are necessarily limited to cases in which a urine culture was ordered. Urine cultures are likely to be ordered in situations in which a patient is very sick, in which case the principle clinical need is to obtain speciation data and antibiotic sensitivities. However, urine cultures might also be ordered in cases in which a patient seems to be clinically stable, to defer treatment pending the culture, with the expectation that the culture will be negative. Focusing analysis on patients for whom a urine culture was ordered therefore likely introduces a bias into our analysis, but we are not sure how strongly (or in what direction) this bias influences our results.

Our present results suggest that future prospective evaluation of the proposed decision rules as a tool to decrease antibiotic overuse is justified and is unlikely to cause harm to nonpregnant adults without high-risk features.

Read or post commentaries in response to this article.

Key words: urinary tract infection; machine learning; antibiotic overuse; decision rule; prediction model; primary care

Submitted March 18, 2022; submitted, revised, August 23, 2022; accepted August 31, 2022.

Funding support: This work was not directly funded but used the REDCap data management platform at the University of Kansas Medical Center, which was supported by a Clinical and Translational Science Awards grant from the National Center for Advancing Translational Sciences (NCATS) awarded to the University of Kansas for Frontiers: University of Kansas Clinical and Translational Science Institute (#UL1TR002366). This work is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or NCATS. This agency had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Supplemental materials

References

1. Medina M, Castillo-Pino E. An introduction to the epidemiology and burden of urinary tract infections. *Ther Adv Urol*. 2019;11:1756287219832172. 10.1177/1756287219832172
2. Mazzulli T. Resistance trends in urinary tract pathogens and impact on management. *J Urol*. 2002;168(4 Pt 2):1720-1722. 10.1097/01.ju.0000028385.10311.c9
3. Butler CC, Rollnick S, Pill R, Maggs-Rapport F, Stott N. Understanding the culture of prescribing: qualitative study of general practitioners' and patients' perceptions of antibiotics for sore throats. *BMJ*. 1998;317(7159):637-642. 10.1136/bmj.317.7159.637
4. Knottnerus BJ, Bindels PJ, Geerlings SE, Moll van Charante EP, ter Riet G. Optimizing the diagnostic work-up of acute uncomplicated urinary tract infections. *BMC Fam Pract*. 2008;9:64. 10.1186/1471-2296-9-64
5. Bent S, Nallamothu BK, Simel DL, Fihn SD, Saint S. Does this woman have an acute uncomplicated urinary tract infection? *JAMA*. 2002;287(20):2701-2710. 10.1001/jama.287.20.2701
6. Heckerling PS, Canaris GJ, Flach SD, Tape TG, Wigton RS, Gerber BS. Predictors of urinary tract infection based on artificial neural networks and genetic algorithms. *Int J Med Inform*. 2007;76(4):289-296. 10.1016/j.ijmedinf.2006.01.005
7. Little P, Turner S, Rumsby K, et al. Developing clinical rules to predict urinary tract infection in primary care settings: sensitivity and specificity of near patient tests (dipsticks) and clinical scores. *Br J Gen Pract*. 2006;56(529):606-612.
8. McIsaac WJ, Moineddin R, Ross S. Validation of a decision aid to assist physicians in reducing unnecessary antibiotic drug use for acute cystitis. *Arch Intern Med*. 2007;167(20):2201-2206. 10.1001/archinte.167.20.2201
9. Wigton RS, Hoellerich VL, Ornato JP, Leu V, Mazzotta LA, Cheng IH. Use of clinical findings in the diagnosis of urinary tract infection in women. *Arch Intern Med*. 1985;145(12):2222-2227.
10. Winkens R, Nelissen-Arets H, Stobberingh E. Validity of the urine dipslide under daily practice conditions. *Fam Pract*. 2003;20(4):410-412. 10.1093/fampra/cmg412
11. Taylor RA, Moore CL, Cheung KH, Brandt C. Predicting urinary tract infections in the emergency department with machine learning. *PLoS One*. 2018;13(3):e0194085. 10.1371/journal.pone.0194085
12. Bishop CM. *Neural Networks for Pattern Recognition*. Oxford University Press; 1995.
13. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California: ACM; 2016:785-794. 10.1145/2939672.2939785
14. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32. 10.1023/A:1010933404324
15. Ho TK. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition. Montreal, Canada: IEEE; 1995;1:278-282. 10.1109/ICDAR.1995.598994
16. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138. 10.1097/EDE.0b013e3181c30fb2
17. Pulivarthi S, Gurram MK. Effectiveness of D-dimer as a screening test for venous thromboembolism: an update. *N Am J Med Sci*. 2014;6(10):491-499. 10.4103/1947-2714.143278