Supplemental materials for

Dhanda G, Asham M, Shanks D, et al. Adaptation and external validation of pathogenic urine culture prediction in primary care using machine learning. *Ann Fam Med*. 2023;21(1):11-18.

Supplemental Appendix 1. Methods

Primary Care data extraction and quality assurance

Model optimization

Model validation

Definition of “high risk” features

Primary Care data extraction and quality assurance

Data for the Primary Care dataset was compiled in two steps. First, we retrieved a list of all adult patients seen in the Family Medicine clinic in 2018 who had at least one visit where both a urinalysis and urine culture were ordered. Records for which a urinalysis or urine culture was ordered – but not completed (N=55) – or for which documentation was not complete (N=1) were excluded. After excluding these records, 472 of the 528 records initially screened were included in the analysis.

Second, data was extracted from medical records manually by 8 licensed physicians (GD, N=123; MA, N=57; DS, N=65; NO, N=65; JH, N=65; MTS, N=63; NTY, N=66; and DJP, N=24). The senior author (DJP) verified, for all records: (1) satisfaction of the inclusion criteria, (2) urine culture pathogenicity annotations, (3) antibiotic prescription annotations. All records flagged by the primary reviewers were also re-reviewed by the senior author. Patients with multiple visits in 2018 in which a urinalysis and urine culture were completed were included only once, using the office visit that occurred latest in the year.

For each record, we extracted data on age, biological sex, urine culture pathogenicity, urinalysis (clarity, protein, glucose, ketones, blood, nitrites, leukocytes), urine microscopy (white blood cells, red blood cells, epithelial cells, bacteria), vitals (temperature, heart rate, respiratory rate, blood pressure), symptoms (dysuria, abdominal pain, subjective fever), history of UTI, higher-risk clinical features (altered mental status, low abdominal pain, flank pain, costovertebral angle tenderness, vomiting, urinary catheter *in situ*, history of renal calculi, immunocompromise), and pregnancy. Urine cultures were considered pathogenic only if they grew more than 100,000 colony forming units (cfu) of an organism that was not a common urogenital or skin flora contaminant. Antibiotics were recorded as prescribed if the prescription was at the time of the office visit (*i.e.*, prior to return of culture). Most urinalyses were performed on a ClinTek Status+ Analyzer (Siemens Medical Solutions, Malvern, PA) and conversions between alternative reporting formats (e.g. ketones 40mg/dL is equivalent to “2+”) was guided by Appendix 10 of the device operations manual.

Model optimization

Training hyperparameters were optimized by grid search with an objective function of maximizing the integral (area under the curve, AUC) of the receiver operating characteristic (ROC) curve.

Model validation

Trained models were (1) internally validated using the Emergency Department 20% hold-out validation set and (2) externally validated on the Primary Care dataset. Model output was evaluated in three ways. First, trained models output a continuous probability that a urine culture will have a pathogenic result. The true positive rate (sensitivity) as a function of the true negative rate (1 - specificity) defines an ROC curve, and we report the area under this curve (ROC-AUC), which summarizes model discriminative performance. ROC curves were compared using DeLong's test. We also computed the scaled Brier score, which is one minus the average squared error of the predicted probability of pathogenicity, normalized against the score of an (uninformative) model that makes uniform predictions.[16]

Second, discrete predictions (pathogenic, not pathogenic) are made by choosing a cutoff probability value: above the cutoff cultures are predicted to be pathogenic, and below the cutoff, cultures are predicted to be nonpathogenic. We characterize the sensitivity (Sen), specificity (Spec), positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+), negative likelihood ratio (LR-) and diagnostic odds ratio (DOR) at (1) the "optimal" cutoff (*i.e.,* the cutoff maximizing the Youden index, sensitivity + specificity - 1), and (2) at a 15% false negative rate (FNR; 85% sensitivity). The significance of the 15% FNR cutoff is that it allows 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.

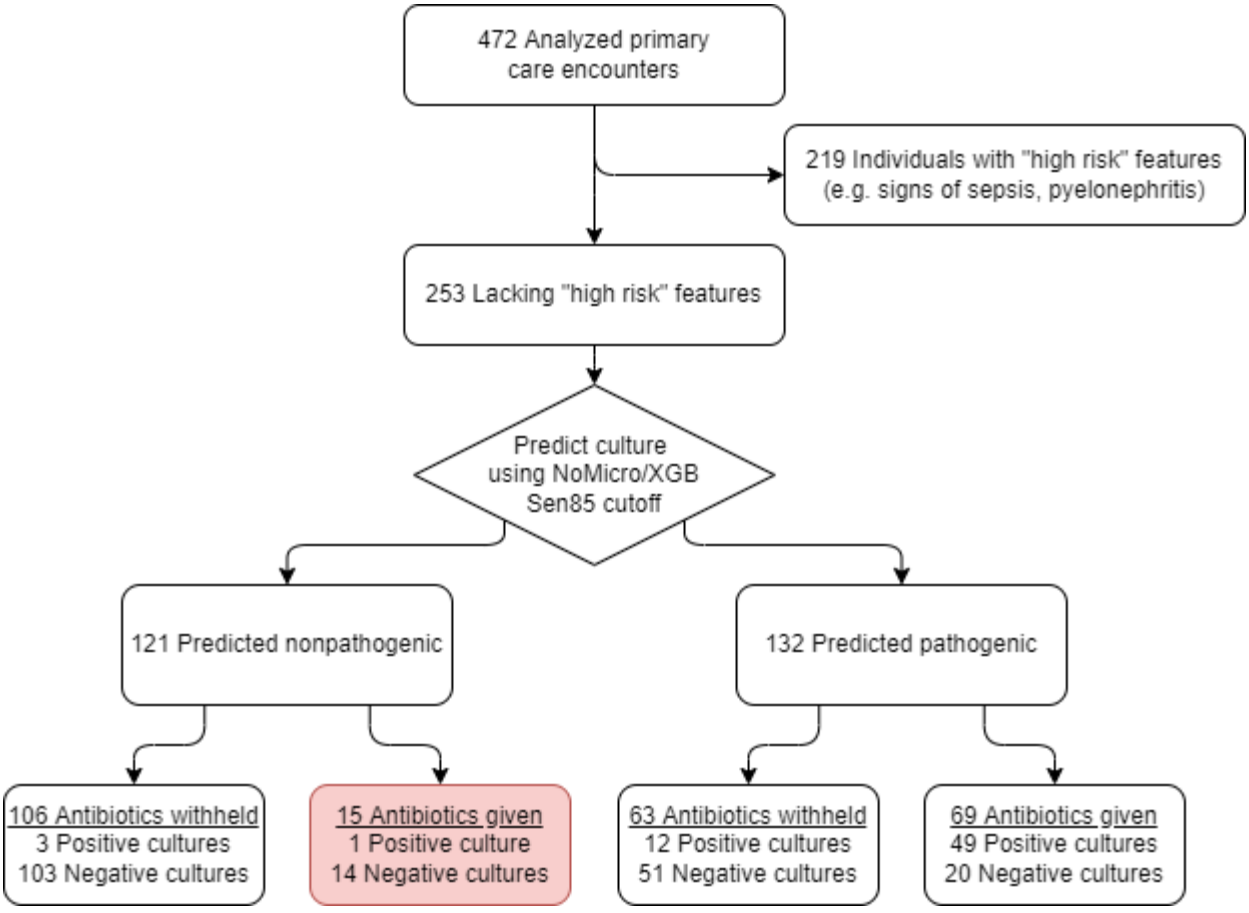
Third, we evaluated model calibration – *e.g.*, a culture with a pathogenicity prediction of 30% should turn out to actually be pathogenic about 30% of the time – by dividing the predictions into 10 equal bins (0-10%, 11-20%, etc., "deciles") and comparing the pathogenicity rate to the mean prediction probability within each bin. Well-calibrated models should approximate a straight line with a slope of 1 and an intercept of 0, with no pattern to the fluctuations above or below the midline. Calibration was first evaluated graphically using decile plots, comparing the mean predicted pathogenicity within each decile to the mean observed pathogenicity within each decile. Second, linear models were fit to the decile plots. The slope and intercept was compared to their expected values (1 and 0, respectively). The coefficient of determination (R^2) was also obtained from these linear fits, which quantitatively describes how much of the variation in the within-decile mean pathogenicity is related to the within-decile predicted pathogenicity. 95% confidence intervals for the parameters described above were estimated using the ROCR R package (for ROC-AUC) or by bootstrapping with 2000 replicates (for all other parameters).

Definition of "high risk" features

Fever greater than 38.0 C (100.4 F), tachycardia greater than 90 beats per minute, tachypnea greater than 21 breaths per minute, systolic blood pressure less than 90 mm Hg, any acute change in mental status or cognition, flank pain, costovertebral angle tenderness, vomiting, presence of a urinary catheter or immunocompromised state.

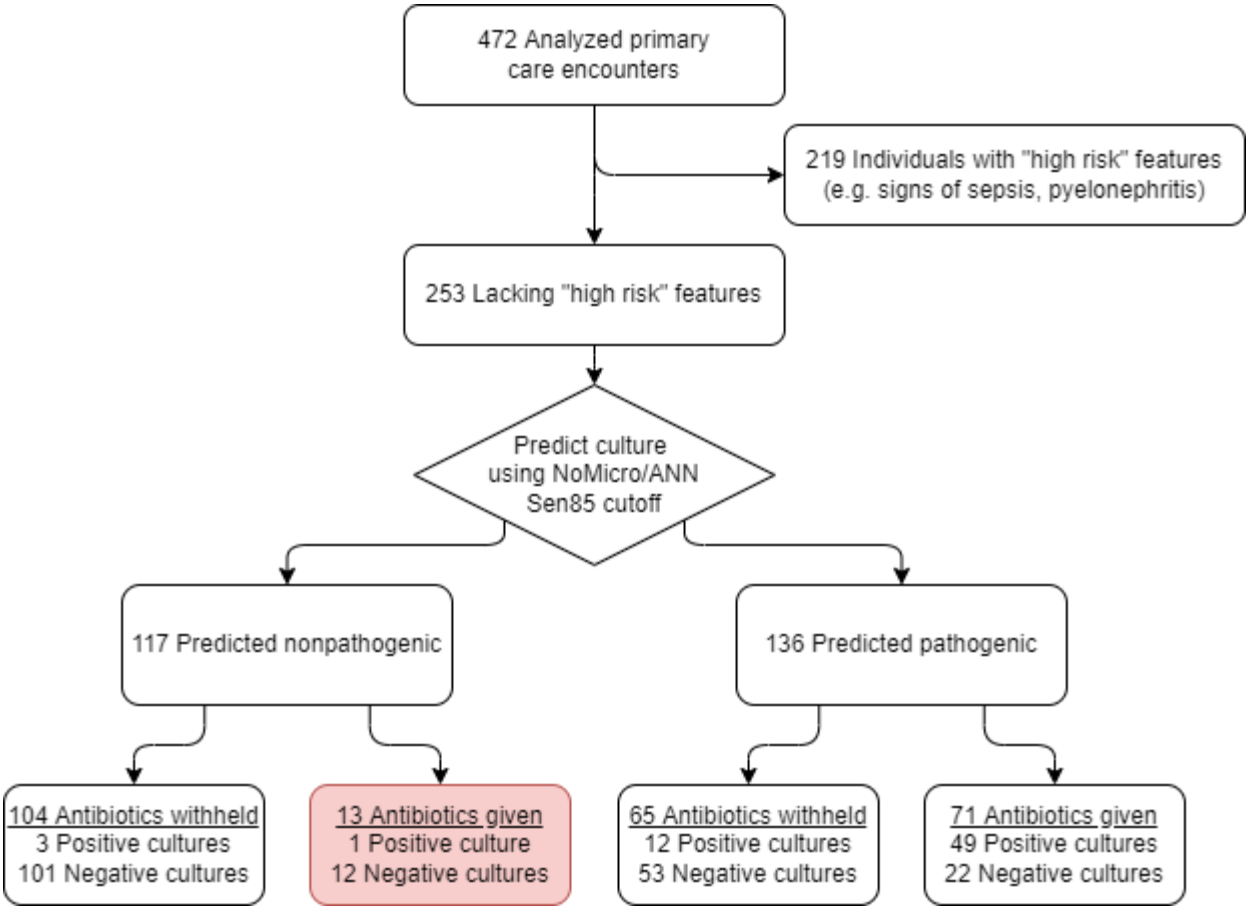
Supplemental Figure 1. Evaluating the potential of NoMicro to reduce antibiotic overuse (Extreme Gradient Boosting)

This is a variation on Figure 1, except that this figure uses the NoMicro/XGBoost classifier instead of the NoMicro/RF classifier. Of 472 primary care encounters, 253 lacked high risk features for progression to serious illness and were stratified using the NoMicro/XGBoost (XGB) 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/XGB model). The cell shaded in red represents cases in which the NoMicro/XGB 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 plausibly reduced by withholding antibiotics in this group.



Supplemental Figure 2. Evaluating the potential of NoMicro to reduce antibiotic overuse (Artificial Neural Networks).

This is a variation on Figure 1, except that this figure uses the NoMicro/ANN classifier instead of the NoMicro/RF classifier. Of 472 primary care encounters, 253 lacked high risk features for progression to serious illness and were stratified using the NoMicro/Artificial Neural Networks (ANN) 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/ANN model). The cell shaded in red represents cases in which the NoMicro/ANN 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 plausibly reduced by withholding antibiotics in this group.



Supplemental Table 1. Model variables stratified by urine culture pathogenicity.

			Emergency Department, No. (%[a])			
	Primary care, No. (%)		Training		Validation	
Feature	Pathogenic	Benign	Pathogenic	Benign	Pathogenic	Benign
Age, years						
18-25	13 (25.5)	38 (74.5)	1742 (21.6)	6335 (78.4)	451 (22.8)	1524 (77.2)
26-35	21 (24.1)	66 (75.9)	1840 (19.5)	7615 (80.5)	468 (19.2)	1968 (80.8)
36-45	15 (17.6)	70 (82.4)	1422 (18.9)	6103 (81.1)	339 (17.6)	1586 (82.4)
46-55	20 (33.9)	39 (66.1)	1816 (18.5)	8009 (81.5)	419 (17.2)	2011 (82.8)
56-65	24 (26.7)	66 (73.3)	1664 (20.2)	6566 (79.8)	438 (20.9)	1659 (79.1)
66-75	22 (32.8)	45 (67.2)	1842 (25.0)	5538 (75.0)	440 (24.0)	1394 (76.0)
>75	13 (39.4)	20 (60.6)	4392 (31.8)	9426 (68.2)	1011 (29.9)	2369 (70.1)
Gender						
Male	11 (17.2)	53 (82.8)	3070 (15.6)	16578 (84.4)	685 (13.9)	4251 (86.1)
Female	117 (28.7)	291 (71.3)	11504 (26.3)	32299 (73.7)	2831 (25.9)	8091 (74.1)
No report	-	-	144 (16.8)	715 (83.2)	50 (22.8)	169 (77.2)
Dysuria						
No	39 (16.5)	198 (83.5)	5101 (17.9)	23318 (82.1)	1187 (16.9)	5828 (83.1)
Yes	75 (37.9)	123 (62.1)	2828 (34.2)	5439 (65.8)	725 (35.5)	1315 (64.5)
No report	14 (37.8)	23 (62.2)	6789 (24.6)	20835 (75.4)	1654 (23.6)	5368 (76.4)
Abd. pain						
No	77 (27.5)	203 (72.5)	5055 (24.1)	15904 (75.9)	1266 (23.9)	4033 (76.1)
Yes	33 (24.3)	103 (75.7)	5549 (18.4)	24658 (81.6)	1354 (17.8)	6238 (82.2)
No report	18 (32.1)	38 (67.9)	4114 (31.3)	9030 (68.7)	946 (29.7)	2240 (70.3)
Hx of UTI						
Yes	68 (36.4)	119 (63.6)	1627 (42.5)	2198 (57.5)	398 (41.3)	566 (58.7)
No	60 (21.1)	225 (78.9)	13091 (21.6)	47394 (78.4)	3168 (21.0)	11945 (79.0)
Blood						
Negative	34 (17.3)	163 (82.7)	5224 (14.7)	30387 (85.3)	1289 (14.4)	7685 (85.6)
Small	53 (33.1)	107 (66.9)	3420 (30.4)	7832 (69.6)	853 (30.3)	1960 (69.7)
Moderate	23 (31.1)	51 (68.9)	2259 (38.0)	3678 (62.0)	509 (35.3)	934 (64.7)
Large	18 (45.0)	22 (55.0)	3685 (34.5)	6990 (65.5)	877 (33.0)	1778 (67.0)
Other	-	1 (100.0)	118 (14.6)	692 (85.4)	38 (20.1)	151 (79.9)
No report	-	-	12 (48.0)	13 (52.0)	-	3 (100.0)
Clarity						
Clear	48 (17.2)	231 (82.8)	3386 (10.4)	29139 (89.6)	846 (10.2)	7416 (89.8)
Not clear	80 (42.1)	110 (57.9)	7316 (36.8)	12552 (63.2)	1741 (35.6)	3150 (64.4)
No report	-	3 (100.0)	4016 (33.7)	7901 (66.3)	979 (33.5)	1945 (66.5)
Glucose						
Negative	122 (28.4)	307 (71.6)	13279 (22.8)	44878 (77.2)	3212 (22.2)	11276 (77.8)
Small	1 (20.0)	4 (80.0)	1088 (25.4)	3188 (74.6)	272 (24.0)	861 (76.0)
Moderate	2 (11.8)	15 (88.2)	98 (23.3)	323 (76.7)	13 (14.8)	75 (85.2)
Large	2 (13.3)	13 (86.7)	235 (16.8)	1160 (83.2)	63 (17.6)	295 (82.4)
Other	1 (16.7)	5 (83.3)	8 (17.8)	37 (82.2)	6 (60.0)	4 (40.0)
No report	-	-	10 (62.5)	6 (37.5)	-	-
Ketones						
Negative	100 (26.1)	283 (73.9)	11875 (22.5)	40974 (77.5)	2865 (21.8)	10305 (78.2)
Small	26 (32.9)	53 (67.1)	2405 (26.6)	6633 (73.4)	597 (25.9)	1712 (74.1)
Moderate	-	4 (100.0)	256 (18.5)	1128 (81.5)	56 (16.8)	277 (83.2)
Large	-	-	137 (15.2)	766 (84.8)	34 (14.7)	197 (85.3)
4+	2 (33.3)	4 (66.7)	6 (54.5)	5 (45.5)	3 (75.0)	1 (25.0)
Other	-	-	28 (26.4)	78 (73.6)	11 (36.7)	19 (63.3)
No report	-	-	11 (57.9)	8 (42.1)	-	-
Leukocytes						

Negative	20 (10.4)	173 (89.6)	2472 (7.0)	32935 (93.0)	615 (6.8)	8384 (93.2)
Small	72 (33.0)	146 (67.0)	6544 (33.8)	12845 (66.2)	1663 (34.5)	3163 (65.5)
Moderate	10 (66.7)	5 (33.3)	2054 (50.5)	2010 (49.5)	486 (48.6)	513 (51.4)
Large	26 (57.8)	19 (42.2)	3610 (67.8)	1711 (32.2)	792 (64.8)	430 (35.2)
Other	-	1 (100.0)	27 (25.7)	78 (74.3)	10 (35.7)	18 (64.3)
No report	-	-	11 (45.8)	13 (54.2)	-	3 (100.0)
Nitrite						
Negative	76 (18.4)	336 (81.6)	10085 (17.4)	47911 (82.6)	2446 (16.8)	12114 (83.2)
Positive	52 (89.7)	6 (10.3)	4592 (74.3)	1587 (25.7)	1108 (74.7)	375 (25.3)
Other	-	2 (100.0)	30 (25.9)	86 (74.1)	12 (35.3)	22 (64.7)
No report	-	-	11 (57.9)	8 (42.1)	-	-
Protein						
Negative	56 (21.5)	204 (78.5)	5523 (15.8)	29487 (84.2)	1377 (15.6)	7449 (84.4)
Small	41 (30.1)	95 (69.9)	7008 (29.9)	16457 (70.1)	1704 (28.9)	4187 (71.1)
Moderate	26 (41.9)	36 (58.1)	1588 (38.0)	2593 (62.0)	354 (35.6)	640 (64.4)
Large	5 (38.5)	8 (61.5)	582 (36.5)	1012 (63.5)	128 (35.9)	229 (64.1)
Other	-	1 (100.0)	4 (11.4)	31 (88.6)	3 (50.0)	3 (50.0)
No report	-	-	13 (52.0)	12 (48.0)	-	3 (100.0)
Micro Bacteria						
None	1 (33.3)	2 (66.7)	629 (9.7)	5828 (90.3)	149 (9.0)	1503 (91.0)
Few	-	9 (100.0)	4167 (21.5)	15241 (78.5)	1016 (21.2)	3779 (78.8)
Moderate	1 (33.3)	2 (66.7)	2743 (40.0)	4113 (60.0)	694 (41.1)	993 (58.9)
Many	1 (100.0)	-	3683 (67.0)	1813 (33.0)	904 (66.3)	460 (33.7)
Marked	2 (40.0)	3 (60.0)	1329 (67.4)	642 (32.6)	316 (66.4)	160 (33.6)
No report	123 (27.3)	328 (72.7)	2167 (9.0)	21955 (91.0)	487 (8.0)	5616 (92.0)
Micro Epi cells						
Negative	3 (15.8)	16 (84.2)	1024 (34.2)	1966 (65.8)	237 (31.2)	522 (68.8)
Small	2 (15.4)	11 (84.6)	8450 (29.8)	19945 (70.2)	2074 (29.0)	5070 (71.0)
Moderate	1 (33.3)	2 (66.7)	1987 (27.6)	5209 (72.4)	472 (27.6)	1236 (72.4)
Large	-	4 (100.0)	743 (22.7)	2529 (77.3)	186 (22.6)	638 (77.4)
Other	-	-	2 (13.3)	13 (86.7)	-	2 (100.0)
No report	122 (28.2)	311 (71.8)	2512 (11.2)	19930 (88.8)	597 (10.6)	5043 (89.4)
Micro WBCs						
Negative	3 (11.5)	23 (88.5)	44 (3.7)	1131 (96.3)	11 (3.9)	274 (96.1)
Small	1 (11.1)	8 (88.9)	3291 (12.9)	22205 (87.1)	806 (12.6)	5602 (87.4)
Moderate	-	3 (100.0)	7013 (44.1)	8887 (55.9)	1711 (43.9)	2189 (56.1)
Large	4 (66.7)	2 (33.3)	3521 (75.3)	1155 (24.7)	846 (74.9)	284 (25.1)
Other	-	-	6 (28.6)	15 (71.4)	-	5 (100.0)
No report	120 (28.0)	308 (72.0)	843 (4.9)	16199 (95.1)	192 (4.4)	4157 (95.6)

[a] Row-wise percentage, not column-wise (*e.g.,* in the primary care dataset 25.5% of cultures *of individuals aged 18-25 years* were pathogenic, as compared to 74.5% of cultures *of individuals aged 18-25 years* were nonpathogenic).

Supplemental Table 2. Linear fit parameters to the per-decile calibration plots.

	Linear Fit Parameters, Parameter (95% CI*)					
	Primary care			Emergency department		
Model	Intercept	Slope	R^2	Intercept	Slope	R^2
NoMicro/XGB†	-0.056 (-0.107--0.000671)	1.03 (0.889-1.16)	0.977 (0.83-0.976)	-0.00157 (-0.0117-0.00854)	0.998 (0.972-1.02)	0.998 (0.99-0.999)
NoMicro/RF‡	0.00267 (-0.0545-0.0668)	0.995 (0.847-1.12)	0.943 (0.77-0.966)	0.0952 (0.0816-0.11)	0.812 (0.782-0.843)	0.995 (0.98-0.997)
NoMicro/ANN§	-0.0653 (-0.115--0.00905)	1.09 (0.942-1.22)	0.973 (0.857-0.975)	0.00741 (-0.00356-0.0183)	0.966 (0.938-0.994)	0.999 (0.991-0.999)
NeedMicro/XGB†	—‖	—‖	—‖	0.00209 (-0.00915-0.014)	0.985 (0.957-1.01)	0.997 (0.988-0.998)

* Estimate and 95% confidence interval across 2,000 stratified (by pathogenicity) bootstrap replicates using the percentage method
† XGB, extreme gradient boosting (XGBoost)
‡ RF, random forests
§ ANN, artificial neural network
‖ The NeedMicro classifier cannot be validated on the primary care dataset because urine microscopy data is not available for almost all records