AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology®

BACTERIOLOGY

Development, Evaluation, and Multisite Deployment of a Machine Learning Decision Tree Algorithm To Optimize Urinalysis Parameters for Predicting Urine Culture Positivity

Jansen N. Seheult,[a] Michelle N. Stram,[b] Lydia Contis,[c] Raymond E. Pontzer,[d] Stephanie Hardy,[e] William Wertz,[e] Carla M. Baxter,[f] Michael Ondras,[e] Paula L. Kip,[f] Graham M. Snyder,[d,g] A. William Pasculle[c,h]

[a]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA
[b]Department of Forensic Medicine, NYU Langone Health, New York, New York, USA
[c]Department of Pathology, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[d]Infection Control and Hospital Epidemiology, UPMC, Pittsburgh, Pennsylvania, USA
[e]Laboratory Service Center, UPMC, Pittsburgh, Pennsylvania, USA
[f]Wolff Center, UPMC, Pittsburgh, Pennsylvania, USA
[g]Division of Infectious Diseases, University of Pittsburgh, Pittsburgh, Pennsylvania, USA
[h]Clinical Microbiology Laboratory, UPMC, Pittsburgh, Pennsylvania, USA

Jansen N. Seheult and Michelle N. Stram contributed equally to the manuscript. Author order was determined alphabetically.

ABSTRACT PittUDT, a recursive partitioning decision tree algorithm for predicting urine culture (UC) positivity based on macroscopic and microscopic urinalysis (UA) parameters, was developed in support of a broader system-wide diagnostic stewardship initiative to increase appropriateness of UC testing. Reflex algorithm training utilized results from 19,511 paired UA and UC cases (26.8% UC positive); the average patient age was 57.4 years, and 70% of samples were from female patients. Receiver operating characteristic (ROC) analysis identified urine white blood cells (WBCs), leukocyte esterase, and bacteria as the best predictors of UC positivity, with areas under the ROC curve of 0.79, 0.78, and 0.77, respectively. Using the held-out test data set (9,773 cases; 26.3% UC positive), the PittUDT algorithm met the prespecified target of a negative predictive value above 90% and resulted in a 30 to 60% total negative proportion (true-negative plus false-negative predictions). These data show that a supervised rule-based machine learning algorithm trained on paired UA and UC data has adequate predictive ability for triaging urine specimens by identifying low-risk urine specimens, which are unlikely to grow pathogenic organisms, with a false-negative proportion under 5%. The decision tree approach also generates human-readable rules that can be easily implemented across multiple hospital sites and settings. Our work demonstrates how a data-driven approach can be used to optimize UA parameters for predicting UC positivity in a reflex protocol, with the intent of improving antimicrobial stewardship and UC utilization, a potential avenue for cost savings.

KEYWORDS reflex protocol, machine learning, receiver operating characteristic, PittUDT algorithm, urine culture

The diagnosis of urinary tract infections (UTIs) from history alone is challenging due to the nonspecific signs and symptoms that can accompany these infections. Various combinations of macroscopic and microscopic urinalysis (UA) parameters in the setting of a clinical history and physical examination are used for making a presumptive diagnosis of UTI (1). When used in patient populations with undifferentiated abdominal pain or even among uncomplicated patients with typical UTI symptoms, the yield of routine urine cultures (UC) is quite low (2–4). Protocols which encourage

Editor Nathan A. Ledeboer, Medical College of Wisconsin

Copyright © 2023 American Society for Microbiology. All Rights Reserved.

Address correspondence to A. William Pasculle, pasculleaw@upmc.edu.

The authors declare no conflict of interest.

Received 14 March 2023
Returned for modification 31 March 2023
Accepted 21 April 2023
Published 25 May 2023

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

more selective use of UC may improve antimicrobial stewardship and UC utilization, a potential avenue for cost savings. Reflex testing refers to an algorithm which uses preliminary screening tests to inform decisions about the need for further tests. It has been applied in many settings in laboratory medicine, including protein electrophoresis and immunofixation, celiac disease serology, and thyroid function tests. Protocols for the implementation of UA with automated reflex to UC have been investigated in several settings, including in urology clinics, other outpatient settings, and in the emergency department (ED) (5, 6). The most commonly used UA parameters are the presence of bacteria, white blood cells (WBCs), nitrites, or leukocyte esterase. Jones et al. have shown that the use of such a reflex protocol can reduce the number of UC performed in an ED setting by up to 39% while missing only 11 of 314 positive UC (3.5%) (7).

Traditional algorithms for UC reflex protocols have relied on receiver operating characteristic (ROC) curve analysis or multivariate (usually logistic) regression analysis using variables which have been selected using an automatic stepwise process or based on expert opinion and clinical judgment. The conventional approach of first performing univariate analyses and including in the multivariable regression model only those variables which meet a predetermined significance level is generally unsuitable for handling complex or non-linear relationships between the predictors and response when they exist. Decision tree algorithms provide an alternate data-driven approach for combining thresholded features for predicting urine culture results from urinalysis parameters. Classification and regression tree algorithms produce an output that is human readable and simple to implement within the current laboratory information system infrastructure. These algorithms also permit differential weighting of false positives and false negatives in order to address class imbalances, i.e., where one class or category is underrepresented in the data set. Differential weighting allows the development of a cost-sensitive algorithm that stresses maximizing the negative predictive value (NPV), which is a primary consideration for a screening technique.

In this study, we trained and validated a recursive partitioning decision tree algorithm for predicting UC positivity based on macroscopic and microscopic UA features, in support of a broader initiative to implement a strong electronic health record (EHR)-based diagnostic stewardship intervention for urine infection testing (UIT) that aims to increase the appropriateness of UC testing. The urine infection testing (UIT) committee focused on maximizing the negative predictive value (NPV) of the UA-to-UC reflex protocol in order for it to be an appropriate screening tool to reduce UC utilization while maintaining prescriber trust. We also describe implementation of this UA reflex protocol under which UC orders are canceled if an accompanying automated UA does not meet prespecified criteria and the method for performance monitoring after protocol implementation.

MATERIALS AND METHODS

Setting and study population. A data set of results for UA and UC performed between 1 January 2017 and 31 December 2017 from adult patients (≥18 years of age) in nonmaternity inpatient and outpatient units at five hospitals of the UPMC (University of Pittsburgh Medical Center) academic health care system was extracted from the electronic health record and laboratory information system (LIS). Only patients with a UA and UC performed within 24 h of each other were included in this analysis. The protocols for algorithm development and the urine infection testing program were ethically reviewed and approved by the University of Pittsburgh's Quality Improvement Committee (projects 1203 and 1722).

Characterization of laboratory procedures and electronic reporting. The UA and UC standard operating procedures (SOPs) were reviewed for the 5 hospital sites to assess for differences in UA and UC parameter reporting, as well as differences in how the urinalysis WAM (Sysmex, Lincolnshire, IL) middleware rules transformed raw counts into binned ordinal values at each site. For the 5 hospital sites, the LIS Sunquest (Sunquest, Tucson, AZ) was queried to determine all order codes used for UA and UC, the names of all urinalysis-associated specimen types, and the LIS test result codes (value identifiers [IDs]) for each of the UA parameters.

Microbiologic methods. All sites performed the microscopic UA using the Sysmex UF-1000i (Sysmex, Lincolnshire, IL), which performs identification and quantification via flow cytometry, and performed the macroscopic UA using the Clinitek Atlas or Clinitek Novus (Clinitek, Ramsey, MN), and manual microscopy was performed when indicated according to each laboratory's SOPs. UC were performed in each hospital's microbiology laboratory using standard methodologies. The UA parameters that were studied included presence of WBCs, bacteria, red blood cells (RBCs), leukocyte esterase, nitrate, specific gravity, pH, character, protein, and blood. UC was considered positive if at least 10,000 CFU/mL of one or more likely urinary

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

pathogens was present, including Gram-negative rods, *Enterococcus* spp., *Staphylococcus saprophyticus*, *Staphylococcus aureus*, group B streptococci, and *Aerococcus*. Urine contaminants, referred to as "mixed skin flora," were considered culture negative; these included *Streptococcus viridans* group, coagulase-negative staphylococci, diphtheroids, *Lactobacillus*, *Gardnerella vaginalis*, *Micrococcus*, and *Bacillus* species. If at least 10,000 CFU/mL of one or more likely urinary pathogens was cultured in the presence of other organisms indicative of "mixed skin flora," then the specimen was considered culture positive.

Data cleansing prior to algorithm development. To generate algorithm rules using the historic data from the 5 sites, the data underwent cleansing to create a single unified data set with values that were comparable across sites. Evaluation of the SOPs, middleware rules, urinalysis-associated specimen types, UA- and UC-related test codes, and the LIS test result codes (value IDs) for each of the UA/UC parameters revealed differences across sites which needed to be comprehensively evaluated, documented, and accounted for before algorithm development. To classify the location type of the patient when the UA/UC was performed, approximate location codes for the 5 hospitals were identified, reviewed, and divided into location categories (Table 1); maternity-specific locations were excluded. All LIS test result codes were identified to retrieve results for all analytes in the UA and UC (for most analytes that make up the UA, each site had its own individual analyte code). Approximately 14 UA and 5 UC test order codes were identified among the 5 sites. Thirty-five values for specimen types were identified and manually reviewed and condensed into the categories represented in the data (Table 1) following data cleansing (e.g., identification of duplicate specimen types with nonmeaningful differences due to the lack of uniform naming practices of each site such as "urine" and "urn" and free text values that did not contribute meaningful additional information such as "scath" free-texted as "urine scath").

Although all sites were using the Sysmex UF-1000i and the Clinitek Atlas or Novus, evaluation of the middleware rules for each site revealed differences in the way raw values were being binned into ordinal values. Using the middleware rules, the range of values for each bin of UA results was identified to establish comparable values between sites. Values for the UA parameters were also condensed into groups sharing an equivalent result where different reporting terms were used (e.g., leukocyte esterase "negative, trace, small, moderate, large" and "negative, trace, P1, P2, P3"). In all cases, the numerical values underlying the ordinal bin names were reviewed to ensure that the same values were being appropriately categorized together. This data cleansing and compiling process was repeated for each analyte for all UA and UC results.

Reflex algorithm derivation. A three-step algorithm was created: macroscopic UA to microscopic UA to UC. The first decision tree (macroscopic-to-microscopic-UA reflex) was trained using parameters obtained from the macroscopic UA: leukocyte esterase, nitrate, specific gravity, pH, character, protein, and blood. All observations that were classified as high risk by the first decision tree were then used to train or validate the second decision tree (microscopic-UA-to-UC reflex), which included parameters obtained from the microscopic UA: WBCs, bacteria, and RBCs. Age, gender, hospital location, and specimen type were not used as predictors since they may not be consistently captured in LIS at all sites in the health care system at the time of specimen collection.

ROC analysis. Receiver operating characteristic (ROC) curves were plotted to compare the accuracies of macroscopic and microscopic UA parameters for predicting positive UC results.

Recursive partitioning algorithm. The "rpart" (8) recursive partitioning algorithm was implemented using "caret" (9) in R version 3.4.2 (Comprehensive R Archive Network) (10) and has been extensively described previously (11). Briefly, the decision tree was built by first finding the single variable which best separated the data into two subgroups known as daughter nodes; after the data were separated, this process was repeated on each subgroup until the subgroups reached a minimum size or until there could be no further improvement. At each node or decision point, sample partitioning was based on the predictor variable which maximized the goodness-of-split, i.e., in which created subgroups were more homogenous or "purer" than the data in the original parent group. Variable importance ranking was based on the Gini index, a measure of node impurity that favors larger partitions compared with other split criteria. Tenfold cross-validation was used to give an internal estimate of misclassification by the decision tree and to avoid overfitting. Balanced accuracy was used as the training metric. The decision tree was pruned, i.e., potentially unnecessary nodes and branches were removed, by specifying the complexity parameter at which the balanced accuracy was highest.

Data from 1 January 2017 to 30 September 2017 were used to train and test the decision tree; the original data were randomly divided into training and test data sets in a 2:1 ratio. The decision trees were trained using the training data set, and then the performance of the optimized decision trees was subsequently determined using the test data set. An annotated script of the learning pipeline that can be used for training similar decision trees using a laboratory's local data is provided in the supplemental material.

Performance targets. The performance targets were established *a priori* based on findings from a single-institution study of 791 consecutive urine samples submitted for aerobic culture at a tertiary care center; in this study, the urine culture positivity rate ($>10^5$ CFU/mL) was 12.9%, and urine bacteria and WBC demonstrated a negative predictive value (NPV) above 97% with a potential reduction in unnecessary urine cultures by up to 55% (12). Since the NPV is inversely proportional to the prevalence of positive urine cultures, a more liberal NPV threshold of 90% was applied to account for the urine culture positivity rate of approximately 25 to 30% at the five hospitals included in this study; by similar logic, a total negative proportion (TNP; percent true negatives plus false negatives as a proportion of all cases) between 30 and 60% was considered acceptable based on an estimated 40 to 50% reduction with 10% tolerance either side, as long as the false-negative proportion (FNP) was below 5% for each site.

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

TABLE 1 Patient demographics, location of specimen collection, and specimen source for all cases by hospital

Characteristic[a]	Hospital 1	Hospital 2	Hospital 3	Hospital 4	Hospital 5	Overall
Hospital bed count	495	380	437	795	520	
No. of cases	3,356	8,342	11,093	9,448	6,122	38,361
Age in yr, mean ± SD (range)	60.1 ± 20.1 (18–116)	41.7 ± 19.7 (18–99)	63.2 ± 21.5 (18–108)	58.6 ± 18.2 (18–100)	65.0 ± 17.4 (18–104)	57.4 ± 21.4 (18–116)
Gender, no. male:no. female (% male:% female)	1,267:2,089 (37.8:62.2)	468:7,874 (5.6:94.4)	3,175:7,918 (28.6:71.4)	4,142:5,306 (43.8:56.2)	2,487:3,635 (40.6:59.4)	11,539:26,822 (30.1:69.9)
Location of specimen collection, *n* (%)						
Emergency department	1,476 (44.0)	2,370 (28.4)	7,175 (64.7)	2,127 (22.5)	1,871 (30.6)	15,019 (39.2)
Intensive care unit	120 (1.4)	248 (2.2)	699 (7.4)	627 (10.2)	1,890 (4.9)	1,880 (4.9)
Inpatient floor unit	851 (10.2)	1,442 (13)	4,016 (42.5)	2,495 (40.8)	9,679 (25.2)	9,634 (25.2)
Outpatient	5,001 (60)	2,228 (20.1)	2,431 (25.7)	1,129 (18.4)	11,598 (30.2)	11,549 (30.3)
Other	0 (0)	0 (0)	175 (1.9)	0 (0)	175 (0.5)	175 (0.5)
Specimen source, *n* (%)						
Clean catch	3,081 (91.8)	7,968 (95.5)	10,781 (97.2)	8,347 (88.4)	5,304 (86.6)	35,481 (92.5)
Foley catheter	92 (1.1)	288 (2.6)	944 (10)	781 (12.8)	2,337 (6.1)	2,330 (6.1)
Loop nephrostomy	2 (0)	0 (0)	5 (0.1)	4 (0.1)	18 (0.1)	18 (0.1)
Nephrostomy	23 (0.3)	4 (0)	38 (0.4)	20 (0.3)	100 (0.3)	100 (0.3)
Straight catheter	248 (3)	13 (0.1)	80 (0.9)	3 (0.1)	344 (0.9)	344 (0.9)
Suprapubic catheter	9 (0.1)	7 (0.1)	34 (0.4)	10 (0.2)	81 (0.2)	81 (0.2)
Time difference between UA and UC, median h (5th to 95th percentiles) (range)	0 (0 to 8.0; −24.0 to 24.0)	0 (−0.7 to 2.0; −23.6 to 24.0)	0 (0 to 0.5; −23.2 to 24.0)	0 (−1.1 to 11.4; −24.0 to 24.0)	0 (−0.1 to 8.3; −23.8 to 24.0)	0 (−0.2 to 5.3; −24.0 to 24.0)

[a]SD, standard deviation; UA, urinalysis; UC, urine culture.

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

Training of decision trees. For the macroscopic-to-microscopic-UA reflex, the training data set comprised 19,511 unique paired UA and UC results with complete data for all predictor variables. Hyperparameter tuning was performed to meet the prespecified performance targets: the minimum node size before a split could be attempted was set as 500 observations, and the minimum bucket size for daughter nodes was set as 250 observations; a false negative was also weighted eight times higher than a false positive in order to optimize the negative predictive value. Repeated 10-fold cross-validation with 10 repeats was used to reduce overfitting.

For the microscopic-UA-to-UC reflex, a separate decision tree was trained using the 11,136 observations from the original training data set above that were predicted as being high risk or positive. The minimum node size before a split could be attempted was set as 500 observations, and the minimum bucket size for daughter nodes was set as 250 observations. A false negative was weighted eight times higher than a false positive in order to optimize the negative predictive value. Repeated 10-fold cross-validation with 10 repeats was used to reduce overfitting.

The two decision trees were then combined into a stepwise UA decision tree algorithm called PittUDT (macroscopic UA to microscopic UA to UC).

Performance evaluation using test data set. The test data set comprised 9,773 unique paired UA and UC results with complete data for all predictor variables. The PittUDT algorithm was applied in a stepwise fashion for classification of observations based on their macroscopic UA and/or microscopic UA results.

A true positive means that the algorithm correctly predicted that the UC was positive, and a true negative means that the algorithm correctly predicted that the UC was negative. A false-positive prediction indicates that the model predicted that the UC was positive when in fact it was negative. A false-negative prediction means that the model predicted that the UC was negative when in fact it was positive.

Sensitivity (recall), specificity, positive predictive value (precision; PPV), and NPV were calculated as previously described (13). Naive accuracy (NA) was calculated as the proportion of all cases that were correctly classified. Balanced accuracy (BA) was calculated as the average of the proportions of each individual class that were correctly classified (i.e., the average of the sensitivity and specificity). The 95% confidence intervals (CI) for test performance metrics were calculated using the exact binomial method.

Standardization of laboratory procedures and electronic reporting and deployment of the reflex protocol. Implementation of the reflex protocol required standardization of UA reporting at the 5 hospital sites, standardization of the middleware rules, generation of LIS rules using the decision tree algorithm, and development of a monthly quality control (QC) program for performance monitoring. To address inconsistent practices among laboratories and to simplify ongoing evaluation, monitoring, and maintenance of the reflex testing procedure, a single SOP was created with a prescribed set of LIS and middleware rules to be enacted prior to implementation of the reflex testing algorithm for clinical use. The SOP was also developed to simplify the process of adopting the algorithm at additional sites in the future. The harmonization process was conducted between January 2018 and December 2020.

Prior to phased implementation of the reflex protocol, rules were developed for the physician order entry module of the hospital information system. Persons ordering UA with reflex to culture were required to designate by means of check boxes that the patient was not pregnant, not immunosuppressed, and not undergoing urologic surgery or that the specimen did not come from outside the bladder (e.g., nephrostomy). These samples were cultured regardless of the UA results. Reflex was not used when a UA independent of culture was ordered for noninfectious reasons. A mechanism was included to permit physicians to request culture of samples which did not meet the UA criteria. Also, a mechanism was created to permit culture independent of UA results for specimens collected in the operating room (OR), the only location where this is permitted. Deployment of the urine infection testing algorithm across the 5 sites began in January 2021 and ended in September 2022.

QC simulation. Data from all samples tested between 1 October 2017 and 31 December 2017 were used to simulate implementation of the validated algorithm at all five sites. A site-specific acceptance sampling plan was developed using a two-stage hypergeometric sampling plan; a web-based calculator for both one-stage and two-stage acceptance sampling plans can be found at https://jnsanalytics.shinyapps.io/ua_reflex_qc_sample_size_calculator (14, 15). The estimated number of urine samples tested during a given 4-week quality control (QC) period per site was calculated based on the average test volume per month at each site in the training data. The tolerance limit for the QC false-negative proportion was set at 20%. In the first stage of the sampling plan, up to four QC false-negative results (m = 4) were permitted per QC period at each site. If no more than four QC false-negative results were observed among the first-stage QC samples (n_1) in a 4-week period at a given site, the process was considered to be in control, and no further QC testing was required during that period. For logistic reasons, the n_1 QC samples were divided evenly among the 4 weeks in each QC period and QC samples were chosen at random from the samples submitted for testing in that week. If there were more than 4 QC false-negative results in the initial QC sample, a second QC sample (n_2) was submitted for testing in that QC period. Conformance was demonstrated if no more than one additional process failure was observed in the second sample. If more than one additional process failure was observed, then QC testing failed for that month and required further investigation into algorithm performance and consideration of expanding the QC sample size to encompass all collected specimens in the upcoming month. The total negative proportion and overall false-negative proportion for all samples at each site were also monitored during the simulation. The tolerance limits for the total negative rate were 30% to 60%. The QC program was simulated 1,000 times, each time drawing a random sample of observations or cases from

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

the QC data period; the proportion of simulations for each site with QC testing failures was determined to evaluate the robustness of the proposed QC program.

Local site verification during phased deployment. Due to the interval between algorithm development and deployment, local site verification to ensure proper functioning of the algorithm at each site was conducted incrementally as the urine infection testing implementation was deployed in a phased manner. An acceptance sampling plan similar to the one described above for QC simulation was designed for local site verification to ensure an acceptably low false-negative proportion. A total of three randomly selected patient specimens predicted by the PittUDT urine infection testing algorithm as low risk, due to either negative macroscopic or negative microscopic analysis described above under "microbiologic methods," were cultured per day. The following guidelines were used to verify local site performance for deployment.

1. Urine specimens were inoculated onto blood, MacConkey, and CNA (colistin and nalidixic acid) agars and incubated in ambient air at 35°C.
2. Cultures were reviewed for clinically significant growth at 24 and 48 hours, postincubation.
    a. Clinically significant growth includes >10,000 CFU/mL of the following organisms:
        I. Any Gram-negative rod
        II. Beta-hemolytic *Streptococcus* species
        III. *Enterococcus* species
        IV. *Staphylococcus aureus*
        V. Yeast or other fungi
    b. Any organism considered a skin contaminant (alpha-hemolytic *Streptococcus*, *Lactobacillus*, or diphtheroids) was considered a negative culture.
    c. Any culture demonstrating clinically significant growth was reviewed by the Infectious Disease team, to determine if it represented a true failure of the urine infection testing process.
3. For the first deployment site, a total of 400 verification specimens were evaluated. If fewer than 7.5% of specimens were culture positive (i.e., false negative), then site verification was considered to have passed. The 7.5% threshold was based on a one-sided one-sample test of proportions to ensure that the NPV was >90%, using an alpha of 0.05 (Stata v17).
4. For the remaining four deployment sites, verification testing was performed for an initial period of 20 consecutive days.
    d. If after 20 days of consecutive testing (~60 specimens), the laboratory identified 0 or 1 positive culture with a likely urinary pathogen, site verification was considered to have passed.
    e. If, during the 20-day window, more than 1 culture grew a significant organism, testing was continued for an additional 20 days (~60 additional specimens).
    f. If, during the second 20-day window, 0, 1, or 2 additional cultures grew a significant organism, verification was considered to have passed.
    g. If, during the second 20-day window, more than 2 cultures grew a significant organism, additional testing and local troubleshooting would be performed under the supervision of the Infectious Disease team and laboratory medical director.

RESULTS

Reflex algorithm derivation. This derivation of the reflex algorithm included 38,361 paired UA and UC cases (Fig. 1); the average patient age was 57.4 years, and 70% of samples were from female patients (Table 1). The majority of specimens originated in the ED (39.2%), outpatient units (30.3%), and inpatient units (25.2%), and more than 98% of specimens were either clean catch (92.5%) or Foley catheter (6.1%) specimens. The difference in sample collection time between the UA and UC specimens was within 5.3 h for 95% of samples. Summary data for results of macroscopic and microscopic UA for all cases are provided in the supplemental material (Tables S1 and S2).

The training data set comprised 19,511 cases, of which 5,223 (26.8%) had a positive UC result. ROC analysis using the training data identified urine WBCs, leukocyte esterase, and bacteria as the best discriminators of UC positivity, with areas under the ROC curve (AUCs) of 0.79, 0.78, and 0.77, respectively; the worst discriminators of UC positivity were specific gravity and pH, with AUCs of 0.48 and 0.51, respectively (Fig. 2). The optimized recursive partitioning decision tree is presented in Fig. 3; the reflex algorithm included use of leukocyte esterase and nitrate for predicting whether microscopic UA should be performed and then, use of WBC and bacteria for predicting UC positivity.

The test data comprised 9,773 cases, of which 2,571 (26.3%) had a positive UC result. Using the held-out test data set, the PittUDT algorithm met the prespecified targets of a 30 to 60% total negative proportion (true-negative plus false-negative

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

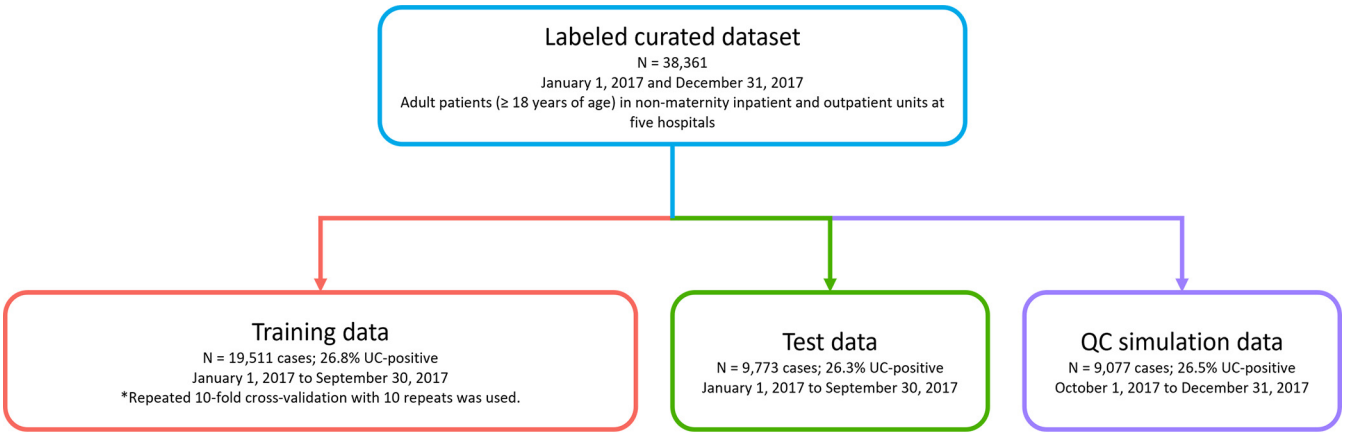

FIG 1 Data flow diagram showing derivation of the training, test, and quality control (QC) simulation data sets.

predictions) with a negative predictive value above 90% and an FNP below 5%. Algorithm performance in the test data set after stratification by age (<60 years versus ≥60 years), gender, and specimen type is shown in Table 2 and Table S3, and performance of the algorithm in the different hospitals and settings (inpatient, outpatient, emergency department, and intensive care unit) is shown in Table 3 and Table S4. The NPV for samples collected from women and ED patients marginally failed to meet the 90% prespecified target, and this was associated with a higher prevalence of positive urine cultures in these two patient groups: 29.3% positive cultures in women compared with 19.4% in men ($P < 0.001$) and 35.9% positive cultures from ED compared with 20.3% positive cultures collected from other sites ($P < 0.001$). There was no significant difference in the NPV (91.2% versus 91.4%, $P = 0.8779$), total negative proportion (53.8% versus 51.8%, $P = 0.07251$), or false-negative proportion (4.7% versus 4.4%, $P = 0.5996$) between paired UA/UC samples collected concurrently (at the same time) and those collected nonconcurrently (Table S5).

QC program simulation. Data from all cases with complete data tested between 1 October 2017 and 31 December 2017 were used to simulate implementation of the

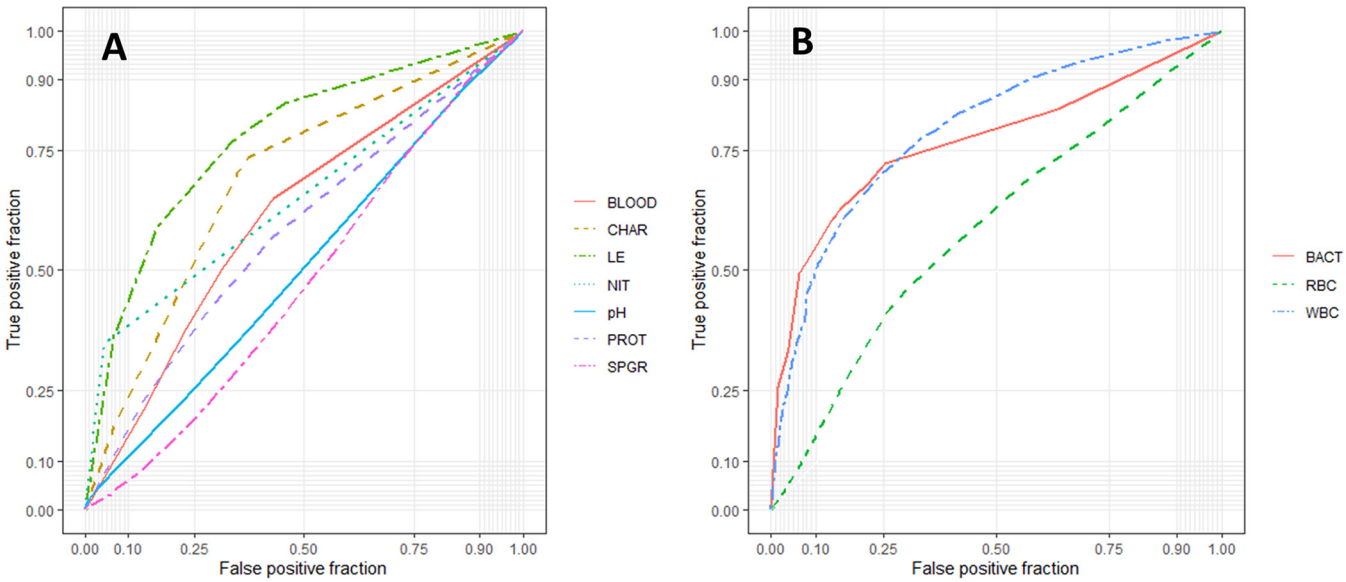

FIG 2 Receiver operating characteristic curves for prediction of urine culture positivity using macroscopic urinalysis parameters (A) and microscopic urinalysis parameters (B). The area under the receiver operating characteristic curve was highest for white blood cells, leukocyte esterase, and bacteria. CHAR, character; LE, leukocyte esterase; NIT, nitrate; PROT, protein; SPGR, specific gravity; BACT, bacteria; RBC, red blood cells; WBC, white blood cells.

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

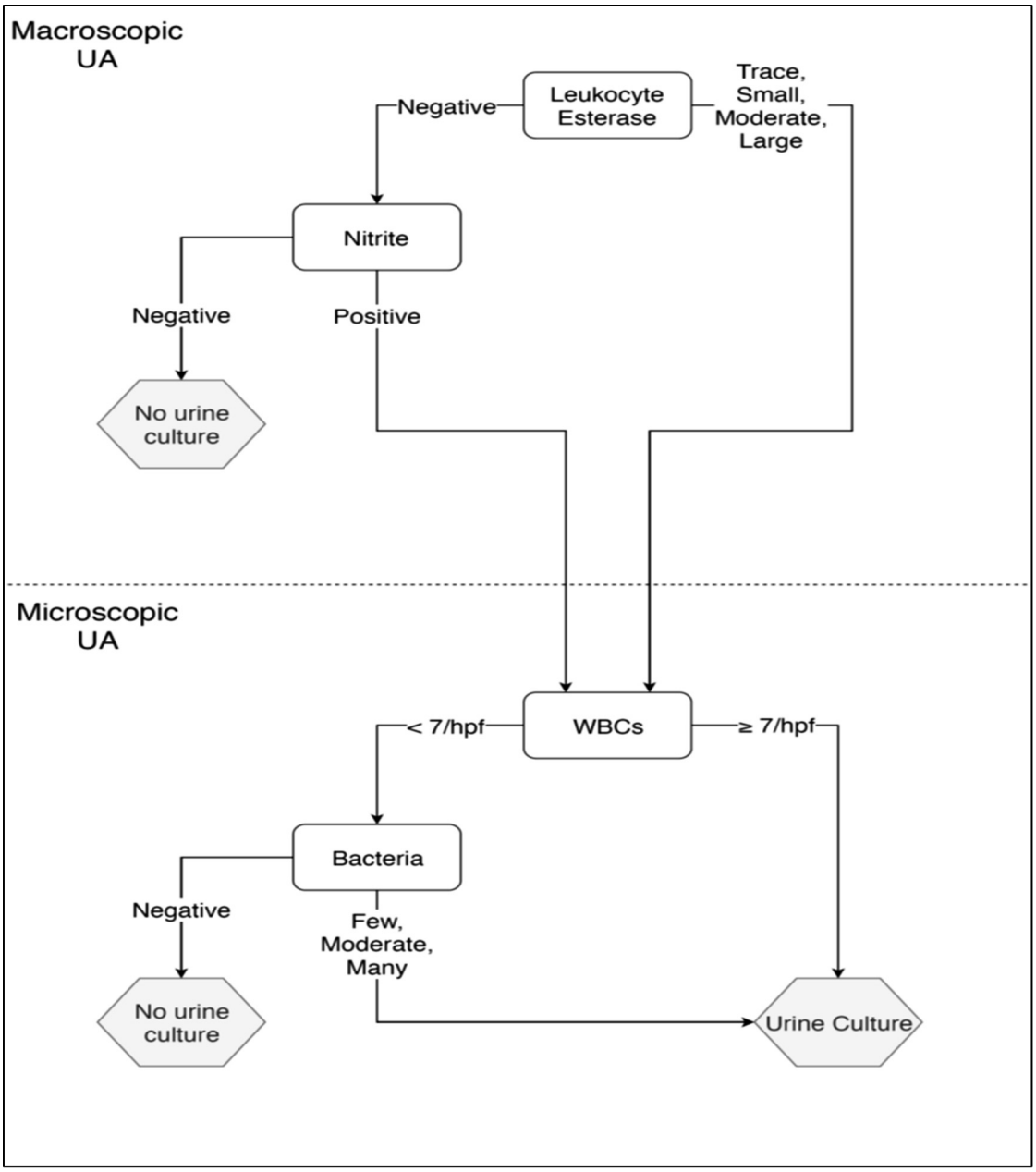

FIG 3 PittUDT algorithm decision tree for prediction of urine culture results using urinalysis parameters.

validated algorithm at all five sites (n = 9,077 cases, of which 2,408, or 26.5%, were UC positive). The results on one simulated quality control (QC) run demonstrating four or fewer QC failures/month across all sites are provided in the supplemental materials (Fig. S1); the actual FNP for each period is also presented. When the QC program was simulated 1,000 times with random QC sampling, the QC testing failure rate was highest for sites with the highest actual FNP for that period (1.5% of simulations for site 1, 4.1% of simulations for site 2, 9.4% of simulations for site 3, 4.5% of simulations for site 4, and 0.4% of simulations for site 5).

Local site verification during phased deployment in real-world use. The performance of the PittUDT urine infection testing algorithm across the 5 hospital sites was examined. A total of 594 randomly selected specimens, which were classified by the algorithm as low risk with the UC not indicated, were cultured. Of these, 589 of the specimens had a negative culture result. Across the 5 hospital sites, the concordance

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

TABLE 2 Summary of test data set decision tree performance by patient demographics and specimen source

	Demographic feature and specimen source						
Performance metric[a]	Overall	Male	Female	Age <60 yr	Age ≥60 yr	Clean catch	Foley catheter
n	9,773	2,961	6,812	4,784	4,989	9,043	592
Requires micro UA, *n* (%)	5,673 (58)	1,383 (46.7)	4,290 (63)	2,486 (52)	3,187 (63.9)	5,191 (57.4)	389 (65.7)
Requires UC, *n* (%)	4,579 (46.9)	1,101 (37.2)	3,478 (51.1)	2,023 (42.3)	2,556 (51.2)	4,158 (46)	336 (56.8)
TNP, %	53.1	62.8	48.9	57.7	48.8	54.0	43.2
Sensitivity, % (95% CI)	82.4 (80.9–83.9)	84.6 (81.4–87.5)	81.8 (80–83.5)	77 (74.3–79.6)	85.8 (83.9–87.4)	81.9 (80.2–83.4)	86.9 (80.6–91.7)
Specificity, % (95% CI)	65.8 (64.7–66.9)	74.2 (72.4–75.9)	61.7 (60.3–63.1)	66.7 (65.2–68.2)	64.9 (63.2–66.5)	66.6 (65.5–67.7)	54.4 (49.6–59.2)
PPV, % (95% CI)	46.3 (44.8–47.7)	44.1 (41.1–47)	47 (45.3–48.7)	37.5 (35.4–39.6)	53.2 (51.3–55.2)	46.2 (44.7–47.8)	41.4 (36.1–46.8)
NPV, % (95% CI)	91.3 (90.5–92.1)	95.3 (94.2–96.2)	89.1 (88–90.1)	91.8 (90.7–92.8)	90.7 (89.5–91.8)	91.3 (90.5–92.1)	91.8 (87.7–94.9)
BA, % (95% CI)	74.1 (73.2–75.1)	79.4 (77.7–81.2)	71.7 (70.6–72.9)	71.9 (70.3–73.4)	75.3 (74.1–76.5)	74.2 (73.3–75.2)	70.6 (67–74.3)

[a]BA, balanced accuracy; CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value; TNP, total predicted negative proportion; micro UA, microscopic urinalysis; UC, urine culture.

of the PittUDT UA results predicted as low risk (i.e. negative predicted UC result) with the actual UC results ranged from 98.4% to 100% (Table 4).

DISCUSSION

The implementation of a multisite conditional UA-reflex-to-UC protocol has the capacity to significantly decrease the number of unnecessary UC which are performed. UA parameters and thresholds used for reflex UC protocols vary widely and have historically been determined based on expert opinion or univariate ROC analysis (12). Our study demonstrates the feasibility and performance of using a supervised machine learning (ML) approach to develop a human-interpretable machine learning model to optimize UA parameters in a reflex protocol that can be deployed across multiple hospital sites and settings within a large academic health care system. Our study also underscores the necessity for a stringent SOP review and robust data cleansing process performed by individuals with a granular understanding of the laboratory processes, middleware rules, and LIS rules, even when the laboratory results are generated at individual laboratories using the same instruments within one health system. Our study shows that a rule-based algorithm trained on UA/UC data has adequate predictive ability for triaging urine specimens by identifying those which are low risk and are unlikely to grow pathogenic organisms, with a false-negative proportion under 5%.

The PittUDT algorithm had similar performance across age, sex, health care setting (inpatient, outpatient, emergency department, and intensive care unit), and specimen type (clean catch, straight catheter, and Foley catheter) strata. The NPV for samples collected from women and ED patients marginally failed to meet the 90% prespecified target, which is likely attributable to the higher prevalence of positive urine cultures in these two patient groups. It is important to note that almost two-thirds of specimens included in algorithm training and evaluation were collected from an acute care setting (ED or inpatient), which may not be reflective of other institutions considering adoption of a similar reflex algorithm. However, one strength of our approach is that the same rules were shown to generalize across health care settings, albeit with the best performance observed in specimens collected in the outpatient setting. The suboptimal performance of the reflex algorithm with nephrostomy specimens may be due to the small sample size in the training and test data sets for this specimen source. However, it is also possible that UA is not an appropriate method of screening urine specimens obtained via suprapubic catheters and loop ileostomies or nephrostomies as patients with indwelling or suprapubic catheters and nephrostomy tubes invariably become carriers of asymptomatic bacteriuria, in which case antibiotic treatment does not have a benefit (16, 17). This highlights an important limitation of any UA-reflex-to-UC protocol, in that it does not prevent testing of asymptomatic bacteriuria. In contrast to other studies on the use of artificial intelligence/machine learning (AI/ML) strategies for reducing unnecessary UC which specifically included special subpopulations of pregnant patients and children, our algorithm is not meant to be applied to these

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

TABLE 3 Summary of test data set PittUDT algorithm decision tree performance by hospital and location of specimen collection[a]

	Hospital and location									
Performance metric	Overall	Hospital 1	Hospital 2	Hospital 3	Hospital 4	Hospital 5	Inpatient	Outpatient	ED	ICU
n	9,773	802	2,195	2,847	2,448	1,481	2,506	3,029	3,760	478
Requires micro UA, *n* (%)	5,673 (58)	629 (78.4)	1,166 (53.1)	1,783 (62.6)	1,190 (48.6)	905 (61.1)	1,246 (49.7)	1,465 (48.4)	2,698 (71.8)	264 (55.2)
Requires UC, *n* (%)	4,579 (46.9)	525 (65.5)	952 (43.4)	1,380 (48.5)	995 (40.6)	727 (49.1)	976 (38.9)	1,132 (37.4)	2,241 (59.6)	230 (48.1)
TNP (%)	53.1	34.5	56.6	51.5	59.4	50.9	61.1	62.6	40.4	51.9
Sensitivity, % (95% CI)	82.4 (80.9–83.9)	91.4 (87.7–94.3)	77.4 (73.3–81.2)	84.5 (81.9–86.8)	77.3 (73.5–80.9)	83.2 (79.1–86.7)	72.3 (68–76.2)	79.5 (76.2–82.6)	87.4 (85.5–89.1)	82.6 (73.3–89.7)
Specificity, % (95% CI)	65.8 (64.7–66.9)	50.4 (45.9–54.9)	65.7 (63.4–67.9)	68.1 (65.9–70.1)	69.1 (66.9–71.1)	63.4 (60.5–66.3)	69 (66.9–71)	74 (72.2–75.8)	56 (54–58)	60.1 (55–65)
PPV, % (95% CI)	46.3 (44.8–47.7)	53 (48.6–57.3)	37.5 (34.4–40.7)	54.9 (52.2–57.5)	39.8 (36.7–42.9)	45.5 (41.9–49.2)	35.8 (32.7–38.9)	45.3 (42.4–48.3)	52.7 (50.6–54.8)	33 (27–39.5)
NPV, % (95% CI)	91.3 (90.5–92.1)	90.6 (86.5–93.8)	91.6 (90–93.1)	90.5 (88.9–92)	92 (90.5–93.4)	91.1 (88.9–93)	91.2 (89.7–92.6)	93 (91.8–94.1)	88.8 (87.1–90.4)	93.5 (89.7–96.3)
BA, % (95% CI)	74.1 (73.2–75.1)	70.9 (68.1–73.7)	71.6 (69.3–73.8)	76.3 (74.7–77.9)	73.2 (71.1–75.3)	73.3 (70.9–75.7)	70.6 (68.3–72.9)	76.8 (74.9–78.6)	71.7 (70.4–73.1)	71.4 (66.6–76.2)

[a]BA, balanced accuracy; CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value; TNP, total predicted negative proportion; micro UA, microscopic urinalysis; UC, urine culture; ICU, intensive care unit.

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

TABLE 4 Local site verification summary by hospital

Facility	Date of reflex implementation (mo/day/yr)	No. of quality control samples	No. of negative cultures on quality control	%
Hospital 3	01/11/2021	400	397	99.3%
Hospital 1	05/10/2021	113	112	99.1%
Hospitals 2 and 4	09/26/2022	60	59	98.4%
Hospital 5	09/26/2022	21	21	100%
Total		594	589	99.2%

special groups (18). There was no uniform reliable mechanism to identify pregnant patients in the LIS, and although maternity-specific locations were excluded, it is unknown which patients in the data set may have been pregnant. Furthermore, the guidance on these special groups is different from that for the general adult population. For example, the U.S. Preventive Services Task Force recommends screening pregnant patients for asymptomatic bacteriuria but not nonpregnant adults (19).

Other studies have developed ML algorithms for specific subpopulations or health care location settings. Taylor et al. applied supervised ML algorithms to predict UTIs in symptomatic patients in the ED, and Burton et al. applied supervised ML algorithms to predict UTIs in both inpatient and outpatient settings with laboratory testing performed in a single clinical microbiology laboratory covering three hospitals and community services (18, 20). In contrast, we developed and deployed an ML algorithm which has broad application (inpatient, outpatient, emergency department, and intensive care unit) and demonstrated that strong clinical performance can be achieved using results from multiple laboratories by employing comprehensive preprocessing steps and domain expertise in understanding what types of laboratory data are truly equivalent and can be combined.

The PittUDT algorithm was created using recursive partitioning to create a human-readable decision tree by an expert group of laboratorians in conjunction with the Health System Infection Control Committee (Fig. 3). This domain expertise was used to determine appropriate weights for false negatives and positives and to determine which metrics should be prioritized (e.g., negative predictive value). Our experience developing this algorithm underscores the importance of the laboratory director's role in ensuring data integrity and equivalency in all of the nuanced decisions that form a foundational part of any ML algorithm implementation.

One concern when implementing a supervised learning algorithm is the ability to monitor for and detect shifts or drifts in algorithm performance over time because of changes in the input data distribution due to, for example, changes in the patient population served or changes in the laboratory methodology, reagents, or instrumentation. This study demonstrates how retrospective data can be used to derive an appropriate acceptance sampling plan based on the hypergeometric distribution that can be applied for ongoing performance monitoring (or quality control) of this and similar algorithms after deployment. Based on the stability of the simulated QC process using 3 months of retrospective data, the UIT team decided to perform performance monitoring on a biannual basis by culturing three low-risk samples per day based on UA results for seven consecutive days. The UIT program is being deployed at additional hospitals in this large multihospital health system, and a detailed review of the quality control program during real-world use will be conducted as more data are generated.

The major limitation of the decision tree algorithm used in this study compared with other supervised learning algorithms is its potential for overfitting due to the high dependence of the algorithm on the training data set that is used in its development. In fact, ensemble methods, such as random forests and gradient boosting, which build many individual trees and then predict based on the consensus of all those trees, are typically preferred since they have a lower risk of overfitting (i.e., lower variance);

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

however, the output from these ensemble methods cannot be easily interpreted by clinicians and laboratory personnel, requiring algorithm orchestration capabilities to be implemented in the LIS or electronic health record (EHR). In this study, minimum parent and daughter node sizes were specified to minimize the overall tree depth, and 10-fold cross-validation was used in order to reduce the risk of overfitting by the decision tree algorithm. In fact, the local site verification data demonstrated that the UIT algorithm continued to meet the prespecified performance target for NPV at the five deployment sites several years after model development and initial evaluation. Other groups have evaluated other, more complex ML methods for creating an algorithm for identifying low-risk UA specimens, which are opaque to the end user. Taylor et al. developed six ML models and concluded that of the ML models they evaluated, the XGBoost algorithm had the best performance (20). In contrast, Burton et al. found that when applied to their cohort, the sensitivity of the XGBoost algorithm was poor (61.7%), and they hypothesized that the difference in performance (sensitivity) could be a consequence of the application of different class weights, as Burton et al. applied class weights that favored high sensitivity (desirable in a screening test) and Taylor et al. did not indicate if weighted classes were used (18).

The development of ML algorithms for use in the clinical setting and the evaluation of their performance cannot be relegated to data scientists working in isolation; it requires clinical expertise and domain knowledge to understand whether appropriate decisions are made in both the selection and implementation of a particular model (e.g., appropriate population selection, equivalence or nonequivalence of laboratory data, understanding the cost of a false-negative or -positive test, and appropriate tuning such as class weighting). Furthermore, unlike other studies where all clinical testing was performed at a single laboratory, the macroscopic and microscopic UA and microbiology testing in our study were performed at multiple laboratory sites, both in the community hospital and in large academic hospital settings. Our study also reveals a cautionary tale for those seeking to exclude laboratorians from these processes and simply aggregate laboratory data generated across multiple sites within a health care system or between health care systems based on the assumption that data generated for the same test, even when performed on the same instrument, should be equivalent. The preprocessing steps and evaluation of the SOPs, middleware, and LIS rules were all integral to developing a sound data set prior to algorithm development.

While the PittUDT algorithm can be fine-tuned for differences in the hospitals, locations within a hospital, and sex, the goal of this study was to create a universal algorithm for all sites utilizing the Sysmex UF-1000i (Sysmex, Lincolnshire, IL) and Clinitek Atlas or Clinitek Novus (Clinitek, Ramsey, MN) urinalysis instruments. As documented in the literature, urine WBCs and bacteria were among the best predictors of positive UC (12, 21). These two parameters are derived from the automated microscopic analysis, which has high reproducibility and accuracy especially when performed by flow cytometry (21). Urine leukocyte esterase, which is a semiquantitative measure of the number of WBCs, was also a good predictor of UC positivity. A universal algorithm based on data-driven parameters with acceptable performance across a diversity of patients is a formidable but attainable achievement which facilitates more rapid implementation and change control in a multisite health care system.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, DOCX file, 1.1 MB.

ACKNOWLEDGMENTS

None of the authors report disclosures.

No funding was obtained to perform this study.

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.

REFERENCES

1. Little P, Turner S, Rumsby K, Warner G, Moore M, Lowes JA, Smith H, Hawke C, Turner D, Leydon GM, Arscott A, Mullee M. 2009. Dipsticks and diagnostic algorithms in urinary tract infection: development and validation, randomised trial, economic analysis, observational cohort and qualitative study. Health Technol Assess 13:1–73. https://doi.org/10.3310/hta13190.
2. Nagurney JT, Brown DF, Chang Y, Sane S, Wang AC, Weiner JB. 2003. Use of diagnostic testing in the emergency department for patients presenting with non-traumatic abdominal pain. J Emerg Med 25:363–371. https://doi.org/10.1016/s0736-4679(03)00237-3.
3. Lammers RL, Gibson S, Kovacs D, Sears W, Strachan G. 2001. Comparison of test characteristics of urine dipstick and urinalysis at various test cutoff points. Ann Emerg Med 38:505–512. https://doi.org/10.1067/mem.2001.119427.
4. McIsaac WJ, Low DE, Biringer A, Pimlott N, Evans M, Glazier R. 2002. The impact of empirical management of acute cystitis on unnecessary antibiotic use. Arch Intern Med 162:600–605. https://doi.org/10.1001/archinte.162.5.600.
5. Fok C, Fitzgerald MP, Turk T, Mueller E, Dalaza L, Schreckenberger P. 2010. Reflex testing of male urine specimens misses few positive cultures may reduce unnecessary testing of normal specimens. Urology 75:74–76. https://doi.org/10.1016/j.urology.2009.08.071.
6. Patel UC, Ismail G, Suda KJ, Sabzwari R, Pacheco SM, Bhoopalam S. 2022. Evaluating the impact of a urinalysis to reflex culture process change in the emergency department at a Veterans Affairs Hospital. Fed Pract 39:76–81. https://doi.org/10.12788/fp.0221.
7. Jones CW, Culbreath KD, Mehrotra A, Gilligan PH. 2014. Reflect urine culture cancellation in the emergency department. J Emerg Med 46:71–76. https://doi.org/10.1016/j.jemermed.2013.08.042.
8. Therneau TM, Atkinson EJ. 2014. rpart: recursive partitioning and regression trees. R package 2014.
9. Kuhn M. 2008. Building predictive models in R using the caret package. J Stat Softw 28:1–26. https://doi.org/10.18637/jss.v028.i05.
10. R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
11. Therneau TM, Atkinson EJ. 2018. An introduction to recursive partitioning using the RPART routines. Mayo Clinic, Rochester, MN.
12. Giesen CD, Greeno AM, Thompson KA, Patel R, Jenkins SM, Lieske JC. 2013. Performance of flow cytometry to screen urine for bacteria and white blood cells prior to urine culture. Clin Biochem 46:810–813. https://doi.org/10.1016/j.clinbiochem.2013.03.005.
13. Weinstein S, Obuchowski NA, Lieber ML. 2005. Clinical evaluation of diagnostic tests. AJR Am J Roentgenol 184:14–19. https://doi.org/10.2214/ajr.184.1.01840014.
14. Venette RC, Moon RD, Hutchison WD. 2002. Strategies and statistics of sampling for rare individuals. Annu Rev Entomol 47:143–174. https://doi.org/10.1146/annurev.ento.47.091201.145147.
15. Chen J, Hong T. 2020. Application of two-stage sampling in sampling inspection, p 1152–1159. *In* International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery.
16. Bonkat G, Widmer AF, Rieken M, van der Merwe A, Braissant O, Muller G, Wyler S, Frei R, Gasser TC, Bachmann A. 2013. Microbial biofilm formation and catheter-associated bacteriuria in patients with suprapubic catheterisation. World J Urol 31:565–571. https://doi.org/10.1007/s00345-012-0930-1.
17. Tenke P, Kovacs B, Bjerklund Johansen TE, Matsumoto T, Tambyah PA, Naber KG. 2008. European and Asian guidelines on management and prevention of catheter-associated urinary tract infections. Int J Antimicrob Agents 31 (Suppl 1):S68–S78. https://doi.org/10.1016/j.ijantimicag.2007.07.033.
18. Burton RJ, Albur M, Eberl M, Cuff SM. 2019. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. BMC Med Inform Decis Mak 19:171. https://doi.org/10.1186/s12911-019-0878-9.
19. US Preventive Services Task Force, Owens DK, Davidson KW, Krist AH, Barry MJ, Cabana M, Caughey AB, Doubeni CA, Epling JW, Jr, Kubik M, Landefeld CS, Mangione CM, Pbert L, Silverstein M, Simon MA, Tseng CW, Wong JB. 2019. Screening for asymptomatic bacteriuria in adults: US Preventive Services Task Force recommendation statement. JAMA 322:1188–1194. https://doi.org/10.1001/jama.2019.13069.
20. Taylor RA, Moore CL, Cheung KH, Brandt C. 2018. Predicting urinary tract infections in the emergency department with machine learning. PLoS One 13:e0194085. https://doi.org/10.1371/journal.pone.0194085.
21. Yusuf E, Van Herendael B, van Schaeren J. 2017. Performance of urinalysis tests and their ability in predicting results of urine cultures: a comparison between automated test strip analyser and flow cytometry in various subpopulations and types of samples. J Clin Pathol 70:631–636.

Downloaded from https://journals.asm.org/journal/jcm on 10 January 2024 by 128.147.28.77.