Risk Factors for Pediatric Sepsis in the Emergency Department
A Machine Learning Pilot Study

Laura Mercurio, MD,∗ Sovijja Pou, BS,† Susan Duffy, MD, MPH,∗‡ and Carsten Eickhoff, PhD§||

Objective: To identify underappreciated sepsis risk factors among children presenting to a pediatric emergency department (ED).

Methods: A retrospective observational study (2017–2019) of children aged 18 years and younger presenting to a pediatric ED at a tertiary care children's hospital with fever, hypotension, or an infectious disease International Classification of Diseases (ICD)-10 diagnosis. Structured patient data including demographics, problem list, and vital signs were extracted for 35,074 qualifying ED encounters. According to the Improving Pediatric Sepsis Outcomes Classification, confirmed by expert review, 191 patients met clinical sepsis criteria. Five machine learning models were trained to predict sepsis/nonsepsis outcomes. Top features enabling model performance (N = 20) were then extracted to identify patient risk factors.

Results: Machine learning methods reached a performance of up to 93% sensitivity and 84% specificity in identifying patients who received a hospital diagnosis of sepsis. A random forest classifier performed the best, followed by a classification and regression tree. Maximum documented heart rate was the top feature in these models, with importance coefficients (ICs) of 0.09 and 0.21, which represent how much an individual feature contributes to the model. Maximum mean arterial pressure was the second most important feature (IC 0.05, 0.13). Immunization status (IC 0.02), age (IC 0.03), and patient zip code (IC 0.02) were also among the top features enabling models to predict sepsis from ED visit data. Stratified analysis revealed changes in the predictive importance of risk factors by race, ethnicity, oncologic history, and insurance status.

Conclusions: Machine learning models trained to identify pediatric sepsis using ED clinical and sociodemographic variables confirmed well-established predictors, including heart rate and mean arterial pressure, and identified underappreciated relationships between sepsis and patient age, immunization status, and demographics.

Key Words: sepsis, machine learning, fever

(*Pediatr Emer Care* 2023;00: 00–00)

Sepsis is the leading cause of inpatient pediatric mortality, accounting for at least 7% to 9% of all pediatric deaths.[1] A significant proportion of pediatric patients who develop sepsis during hospitalization present to the emergency department (ED) for their initial care, with 100,000 children in the United States diagnosed annually with severe sepsis.[2,3] The costs of pediatric sepsis are significant not only because of acute care, but also because of the lifelong disability experienced by up to 40% of survivors.[1,4–6] Timely sepsis recognition with prompt antibiotic and fluid administration are associated with improved outcomes,[2,7,8] but physicians continue to struggle in distinguishing clinical signs of organ dysfunction in children whose abnormal vital signs and examination stem from evolving sepsis as opposed to pain, fever, or common respiratory illness.[2] Furthermore, the formal definition of sepsis in pediatric medicine continues to evolve.[1,9,10] Building on work by Goldstein et al, the Children's Hospital Association's Improving Pediatric Sepsis Outcomes (IPSO) Collaborative expert panel has developed a sepsis definition based on "intention-to-treat" parameters, producing the following classification: patients with infection, patients "at risk" for sepsis, patients with sepsis, and those with "critical sepsis", including shock. Patients were sorted into these categories according to clinical indicators such as fluid administration, vasopressor use, and serologic signs of ongoing organ dysfunction after initial treatment.[11]

Machine learning—a subset of artificial intelligence—is an innovative method for risk factor identification because it provides researchers with 2 distinct advantages: the ability to rapidly process large data sets and construct models that account for complex nonlinear relationships between patient features, time-course, and outcome. Recently, researchers have applied machine learning to predict pediatric sepsis onset and associated organ dysfunction, but these studies have occurred mainly in severely ill patients in the intensive care unit setting[12–14] because of the sparsity of data in other clinical environments. Scott et al[15] were among the first to apply machine learning to an ED data set to formulate a predictive model identifying septic shock among children meeting clinical sepsis criteria, whereas Goto et al[16] applied machine learning to a larger pediatric emergency department (PED) population to assess traditional triage systems and predict those patients ultimately requiring critical care. Very few studies have evaluated the full breadth of structured clinical and sociodemographic variables available from ED encounters to determine their significance in identifying patient characteristics predicting pediatric sepsis.

In this study, we apply machine learning models to patients aged 0 to 18 years presenting to a PED with fever, hypotension, or infectious complaint. Our primary aim was to model those aspects of children's health status and physiologic state available during ED care that were most associated with a sepsis diagnosis during hospitalization. We hypothesized that broad consideration of structured ED clinical data elements, including those associated with underlying health status, would identify prevalent patient sepsis risk factors. Our secondary aim was to apply machine learning to identify additional sociodemographic factors that may contribute to health disparities in patients diagnosed with sepsis.

METHODS

We performed a single-center derivation study to build and internally validate a machine learning model predicting an

From the *Section of Pediatric Emergency Medicine, Department of Emergency Medicine, Alpert Medical School of Brown University, Providence, RI.; †Alpert Medical School of Brown University, Providence, RI.; ‡Department of Pediatrics, Alpert Medical School of Brown University, Providence, RI.; §Brown Center for Biomedical Informatics, Brown University, Providence, RI.; and ||Department of Computer Science, Brown University, Providence, RI.

Disclosure: The authors declare no conflict of interest.

Reprints: Laura Mercurio, MD, 55 Claverick St, 2nd Floor, Providence, RI 02903 (e-mail: laura.mercurio@brownphysicians.org).

Grants: This study was partially funded by the Brown Emergency Medicine Summer Research Assistantship Program (SRA-EM).

Meetings/Conferences: Alpert Medical School at Brown University Spring Research Symposium (2019); Brown University's Unconference on Computational Intelligence & Applications (2020)

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.pec-online.com).

ISSN: 0749-5161

encounter-level diagnosis of sepsis among children presenting to a PED. We then examined the model's contributing features with the goal of identifying pediatric sepsis risk factors. This study was approved by the institutional review board of the Lifespan Healthcare system.

Population Selection

The study population included patients aged 0 to 18 years presenting to a pediatric tertiary care children's hospital ED from January 1, 2017 through January 1, 2019 (Fig. 1). Those individuals with a chief complaint of fever, documented fever greater than 100.4°F, age-adjusted hypotension, or an infectious disease diagnosis based on International Statistical Classification of Diseases and Related Health Problems (ICD-10) (N = 35,047, Appendix B, http://links.lww.com/PEC/B54) were included. Those patients admitted directly to an inpatient service from an outpatient clinic, or transferred directly from an outside hospital, were excluded due to data access limitations. Patient encounters meeting inclusion criteria were then divided into 2 main outcome categories based on criteria set by the IPSO Collaborative (see Outcome Classification section).

Feature Extraction

All readily available, structured electronic health record data in EPIC (Epic Systems, Verona, WI) were extracted for each qualifying patient encounter using structured query language (Fig. 1). The initial data set included vital signs, pediatric clinical scoring tools (eg, pediatric early warning score), intake/output, sociodemographic information, patient problem list, physician orders (eg, medications, fluids), and laboratory studies occurring in at least 1% of encounters.

Final feature categories included the following 76 variables (see Appendix D, http://links.lww.com/PEC/B54 for full listing):

- Sociodemographic information (N = 10) – eg, insurance, immunization status, race, ethnicity. Patients are asked to self-identify race and ethnicity at the time of registration.
- Patient Baseline Problem List (N = 21) – aggregated into organ system-based clinical classification software (CCS) code categories from 3496 unique raw ICD-10 codes
- Emergency Department Nursing Flow Sheets/Vital Signs (N = 45) – eg, heart rate (HR), blood pressure (BP), pulse oximetry, capillary refill measured while in the ED

Feature Selection andMissing Data

After initial feature extraction, nursing flow sheets and vital signs were limited only to those obtained during the patients' time in the ED. The study team then removed those features representing actions taken by the ED provider *in response* to the patient's condition, such as laboratory studies, clinical scoring systems, medication orders, imaging, and procedures. These data were removed to examine the relationship between sepsis and patients' sociodemographic factors, baseline vital signs and medical problems, and physiologic responses to ED treatments rather than to specific interventions or treatments.

Diagnostic codes were also obtained for each patient encounter. All available sociodemographic features and patient problem lists were considered for inclusion. These features included all information available at the time of data extraction; these could include problems added after the qualifying encounter. The ICD-10 problem list was obtained for each patient visit, and these features were consolidated into the 21 categories of the CCS, which are organized by organ system[17] (Appendix C, http://links.lww.com/PEC/B54).

We included those vital sign features occurring in at least 4% of all patient encounters. Vital sign values included the first, last, minimum, mean, and maximum values for HR, diastolic BP, systolic BP, mean arterial pressure (MAP), respiratory rate, pulse oximetry, and body temperature. Mean arterial pressure was calculated from extracted data using the following formula: (⅔ × diastolic BP) + (⅓ × systolic BP). Missing vital sign values were imputed using standard vital sign ranges for age according to the pediatric advanced life support algorithm published by the American Heart

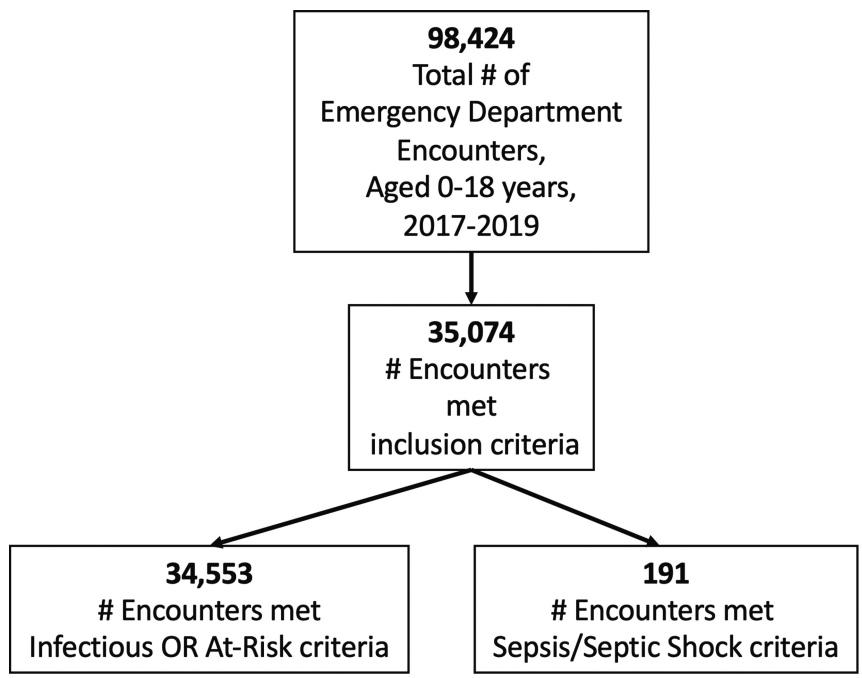

FIGURE 1. Patient encounter inclusion flow diagram.

Association.[18] Blood pressure was the most common missing value at 1.6% of a representative subsample; other vital signs had less than 0.4% missing values.

For nursing clinical scoring systems—eg, pediatric respiratory score, Glasgow coma score, and pediatric asthma severity scores—missing values were coded as a 0, whereas those with documented scores included those values greater than 0. A small number of patient encounters included nursing evaluation checklist values for "tachypnea", "tachycardia", "altered capillary refill", and "altered mental status" documented in the chart—if these were present, they were coded as a 1, otherwise all values were coded as 0. Additional mandatory nursing documentation such as central line–associated bloodstream infection evaluation and total parenteral nutrition were coded as binary variables, with 1 representing the evaluation took place and 0 if it did not occur.

On ED arrival, the triage nurse documents a patient's current immunization status; this information is entered into the electronic health record as a structured variable. Immunization status, "UTD Immunization", was therefore categorized as a ternary variable in which 0 represented not up-to-date, 1 represented up-to-date, and 2 represented unknown/not documented vaccination status. Health insurance was coded as a ternary variable using public, private, and unspecified categories. All patients had a documented home zip code. Other sociodemographic and baseline problem list data were coded as binary variables; for example, if the patients did not have a particular problem, the value was recorded as a 0. Data points were then transformed into a Gaussian distribution with 0 mean and unit variance.

Outcome Classification

All encounters meeting IPSO sepsis treatment criteria for sepsis or critical sepsis underwent chart review by a pediatric sepsis expert to confirm the outcome of interest. Expert outcome classifications were compared against available ICD-10 diagnoses and underwent a secondary review by another clinical expert if the outcome category was unclear (Appendix A, http://links.lww.com/PEC/B54). Patient encounters were assigned to 1 of 4 IPSO expert consensus outcome categories, which were combined to create a binary sepsis versus nonsepsis outcome for ED patients requiring admission:

- Infection or At-Risk (N = 34,883 ED encounters)
  a. Infection: patients with an infectious diagnosis, without meeting other criteria.
  b. At Risk for Sepsis: patients with infection in which a blood culture is obtained and antibiotics are given within 24 hours. They could receive up to 20 mL/kg of isotonic fluid within 6 hours of antibiotics.
- Sepsis or Critical Sepsis/Septic Shock (N = 191 ED encounters)
  a. Sepsis: patients with infection with initial organ dysfunction that *responds* to ED treatment. They received treatment with intravenous antibiotics, at least two 20 mL/kg isotonic fluid boluses, within 6 hours of presentation.
  b. Severe Sepsis/Septic Shock: patients presenting with evidence of infection *with organ dysfunction not responsive to initial treatments* as described, requiring and ongoing resuscitation greater than two 20 mL/kg intravenous isotonic fluid boluses, vasopressors, and intensive care management.

Analysis

Analyses were performed using Python 3.7.3[19] (Python Software Corp, Wilmington, DE). All features first underwent χ^2 analysis to determine their independent association with the outcome of interest (Fig. 2). Candidate models representing popular machine learning approaches were then constructed using features as listed previously (N = 76). These models included Gaussian naive Bayes, support vector machines (SVM), logistic regression (LR), classification and regression trees (CART), and random forest classifiers (RF). We followed common machine learning practices, splitting the available data into nonoverlapping 80%/20% training/testing subsets. Within the 80% training portion, we performed cross-validation to fit model parameters and tune hyperparameters. Given significant outcome class imbalance, the synthetic minority oversampling technique was applied to the data. Final model evaluation was then performed on the withheld test set.

We evaluated each model's ability to correctly identify patients who developed sepsis or critical sepsis/septic shock during their ED visit or inpatient hospitalization using the following performance metrics: recall (sensitivity), specificity, precision (positive predictive value), the area under the receiver operating curve, and F-1 scores (the harmonic mean of precision and recall) (Table 1). To account for the significant outcome class imbalance, we also reported the area under the precision-recall curve (AU-PRC). The trainable model parameters were adjusted via backpropagation of class-balanced cross-entropy loss. Model hyperparameters were tuned to optimize for sensitivity using the scikit-learn software library (Appendix D, http://links.lww.com/PEC/B54).[19,20] Those features determined to be duplicative, or a direct proxy of patient disposition, were removed to reduce collinearity.

We extracted the top quartile of features (N = 20) contributing to the best models' performance based on the "importance coefficient value" (IC) assigned to each feature (see Appendix E for definition, http://links.lww.com/PEC/B54). This coefficient value reflects how much the model depended on each individual feature to distinguish between septic and nonseptic patients.

Stratified analyses were then performed by age, primary language, assigned sex, race, ethnicity, oncologic history, and insurance type to evaluate for confounding between identified sepsis

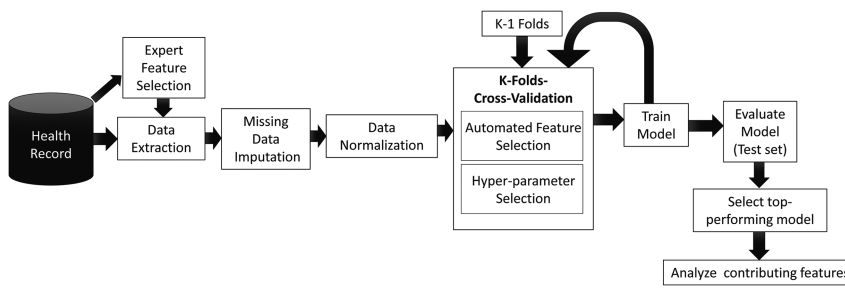

FIGURE 2. Data collection and machine learning model workflow.

TABLE 1. Comparison of Machine Learning Model Performance in Classifying Patients With Sepsis

Performance Characteristic	Random Forest	Classification and Regression Tree	Logistic Regression	Support Vector Maximization	Gaussian Naïve Bayes
Sensitivity/recall	0.93	0.85	0.76	0.70	0.37
Specificity	0.84	0.70	0.88	0.92	0.94
Precision	0.04	0.02	0.04	0.05	0.04
F-1 score	0.07	0.04	0.08	0.10	0.07
Area under the receiver operating curve	0.81	0.77	0.82	0.81	0.65
AU-PRC	0.48	0.43	0.40	0.37	0.11

predictors and sociodemographic factors. First, subpopulations were created according to the categories defined in Table 2. The best performing model (RF) was trained on each subpopulation, and the top quartile of features was extracted and qualitatively compared. For example, the patient population was divided into 2 categories according to self-reported ethnicity, and the model was trained separately for patients identifying as Hispanic and non-Hispanic. We then qualitatively compared top contributing features for the resulting models to identify changes in patient risk factors based on membership in a particular subpopulation.

TABLE 2. Patient Demographics, by Binary Sepsis Outcome Category (N = 35,074)

Characteristic	At Risk for Sepsis or Infection (N = 34,883) N (%)	Sepsis/Septic Shock (N = 191) N (%)	Overall (N = 35,074) N (%)
Age category (y)			
Neonate (0.0–0.1)	685 (2.0)	12 (6.3)	697 (2.0)
Infant (0.11–1.0)	7132 (20)	20 (10)	7152 (20)
Toddler: (1.1–3.0)	10,643 (31)	33 (17)	10,676 (30)
Preschool: (3.1–5.0)	4868 (14)	25 (13)	4893 (14)
School-age: (5.1–12.0)	8189 (23)	56 (29)	8245 (24)
Adolescent: (12.1–18.0)	3366 (9.6)	45 (24)	3411 (9.7)
Assigned sex			
Male	18,346 (53)	92 (48)	18,438 (53)
Female	16,536 (47)	99 (52)	16,635 (47)
Not reported	1	0	1
Race*			
White	13,530 (39)	96 (50)	13,777 (39)
Black	4475 (13)	18 (9.4)	4493 (13)
Asian American	675 (1.9)	5 (2.6)	680 (1.9)
Native Hawaiian	33 (<0.1)	0	33 (<0.1)
American Indian	46 (0.1)	2 (0.1)	48 (0.1)
Other	16,089 (46)	70 (37)	16,159 (46)
Ethnicity*			
Hispanic	15,957 (46)	68 (36)	16,025 (46)
Not Hispanic	18,926 (54)	123 (64)	19,049 (54)
Health insurance			
Public	26,224 (75)	128 (67)	26,352 (75)
Private	7327 (21)	61 (32)	7388 (21)
Unspecified	1332 (3.8)	2 (1)	1334 (3.8)
Oncologic status			
Cancer diagnosis	394 (1.1)	19 (10)	413 (1.2)
No cancer diagnosis	34,489 (99)	172 (90)	34,661 (99)
Immunization Status*			
Up-to-date	29,282 (84)	125 (65)	29,407 (84)
Not up-to-date	968 (2.8)	7 (3.7)	975 (2.8)
Unreported	4633 (13)	59 (31)	4692 (13)

*As self-reported by patient or caregiver at time of registration.

RESULTS

Patient Population

During the 2-year study period, 98,424 patients aged 0 to 18 years sought care in the PED. The final study population included 35,074 patient encounters meeting inclusion criteria, with 191 patients meeting critical sepsis/septic shock criteria (Fig. 1). Most (67%) patients were aged younger than 5 years (Table 2). Sepsis cases (those meeting criteria for sepsis or critical sepsis/septic shock) occurred in 0.54% of the study population. The prevalence of sepsis was increased among adolescents (aged 12–18 years, 1.3%) compared with the school-aged (aged 5–11 years, 0.6%) and the under-5-year-old age groups (0.38%).

There was an increased prevalence of sepsis/critical sepsis among those with undocumented immunization status (1.25%) and not up-to-date immunizations (0.7%), compared with those with up-to-date childhood vaccinations (0.4%) documented in the ED. Finally, 1.1% of the population had an oncologic diagnosis, and the prevalence of sepsis or critical sepsis among this group was 4.6% compared with 0.4% among those without cancer.

Descriptive Analysis (χ^2)

The χ^2 analysis was performed to identify those patient features independently associated with a clinical diagnosis of sepsis/critical sepsis among hospitalized patients admitted through the ED. We examined the top quartile (N = 20) of associated features (Fig. 3). In general, patients' preexisting health conditions—as defined by CCS codes—were independently associated with sepsis/critical sepsis (Fig. 2), with "MAL"—representing congenital abnormalities—as the leading condition (Appendix C, http://links.lww.com/PEC/B54). Emergency department vital sign abnormalities were also present among the top quartile of features of patients with sepsis/critical sepsis including tachypnea, tachycardia, as well as features reflecting clinical assessment and demonstrating signs of organ dysfunction—eg, altered capillary refill (Fig. 3).

Model Performance

Given the high morbidity and mortality associated with failure to recognize subtle clinical signs of sepsis, we paid particular attention to model sensitivity, minimizing false-negative classification. The RF model performed the best (Fig. 4), followed by the CART model (Fig. 5). The RF model demonstrated a sensitivity of 93%, a specificity of 84%, and an F-1 score of 0.07 (ref: 0–1) (Table 1). The CART model returned a sensitivity of 85%, a specificity of 70%, and an F-1 score of 0.04 (ref: 0–1). The SVM model generated a higher F-1 score (0.10) but was limited by a low sensitivity (70%). All models exhibited a high negative predictive value and a low AU-PRC due to outcome class imbalance.

Identified Patient Risk Factors

Once the top-performing models were identified, we performed a qualitative analysis of clinical indicators present during the ED stay and their contribution to each model's ability to predict sepsis/critical sepsis in admitted patients. We extracted the top quartile of features (N = 20) from the 2 best-performing models, RF and CART, and reviewed each set. Each feature was assigned an IC, which reflects how much an individual feature contributed to the model's prediction of patient outcome. Regardless of treatment, maximum documented HR was the top contributing feature in both RF and CART models, with ICs of 0.09 and 0.21, respectively. Both models also ranked maximum MAP as the second most important contributing feature (Fig. 4). Finally, minimum systolic BP was a significant contributor in these models (IC 0.04, 0.06).

Among the measured sociodemographic features, patient age was identified as one of the top contributing features (IC 0.03) enabling model classification of septic patients. The LR was used to determine the directionality of this feature, revealing *increasing* age to be associated with an increased likelihood of a sepsis diagnosis.

Documented immunization status was also identified as a top feature in the RF (IC 0.02). This factor did not show statistical significance in the independent χ^2 correlation analysis (Fig. 3) but became a top contributing feature to the RF, LR, and SVM machine learning models (Fig. 4). Additional key contributing features included ED vital sign parameters such as mean, maximum, first, and last recorded HR, BP, respiratory rate, and oxygenation level.

Finally, patient home zip code was the sixth most significant feature enabling the CART model to predict sepsis/critical sepsis patients (IC 0.02, Fig. 5). This model also incorporated multiple features from the patient problem list.

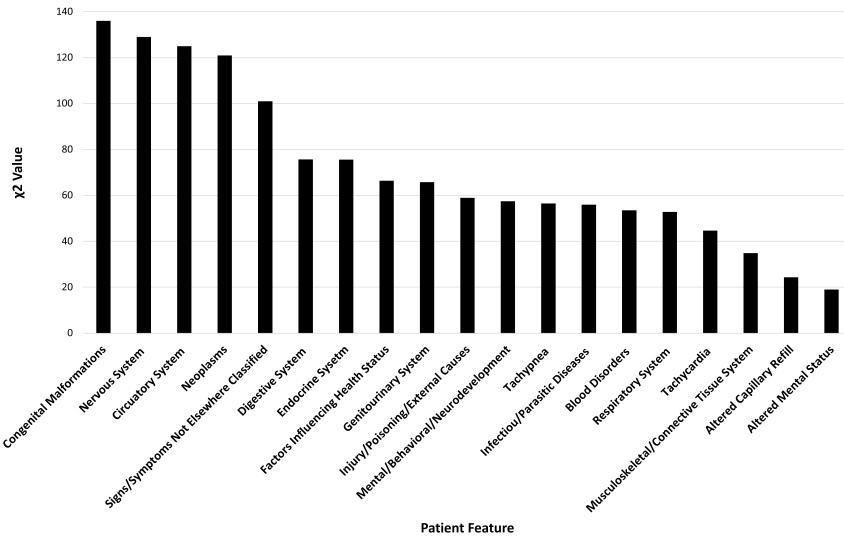

FIGURE 3. Top quartile of patient features independently associated with a sepsis diagnosis on χ^2 analysis (N = 35,074). The larger the χ^2 value, the greater the independent association of that variable with an encounter-level diagnosis of sepsis. Documented, baseline problem lists were grouped into ICD-10 CCS categories, which are organized roughly by organ system, including additional categories influencing health status.

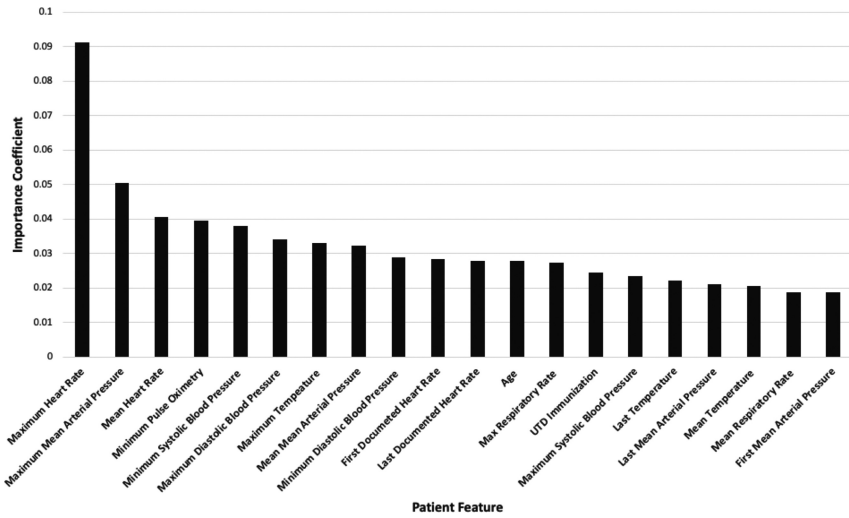

FIGURE 4. Random forest model: top quartile of patient features predicting sepsis (N = 35,074). First indicates first documented in the ED; last, last documented in the ED.

Subpopulation Analyses

When stratified by race, the CCS category broadly represented health care access; FAC (factors influencing health status and contact with health services) ranked in the top quartile (16th), with an IC of 0.02 among non-White patients. This problem list category was *not* listed among the top contributing features for White patients.

Among patients with oncology diagnoses, maximum HR remained the top contributing feature, with 2 problem list categories rising into the top quartile of features: BLD (diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism) and GEN (diseases of the genitourinary system). Among patients without oncologic diagnoses, the top 3 contributing features included: MAP measurements for age, oxygen saturations, and HR.

Patient home zip code was a significant contributing feature among privately insured patients only (IC 0.03). The CCS code for social determinants of health—FAC—was also identified as a top-quartile feature among privately insured patients (IC 0.02).

There were no significant differences in feature importance when stratified analysis was performed by ethnicity or assigned sex.

DISCUSSION

In this study, we applied machine learning to identify the contribution of vital signs, medical problems, and sociodemographic factors present during the PED encounter in predicting pediatric sepsis/critical sepsis diagnosis or treatment during hospitalization that may be difficult to appreciate using conventional correlation analysis alone. Top-performing models confirmed well-established predictors of serious illness such as maximum HR, ongoing HR (HR_first, HR_last), changes in MAP (MAP_max, MAP_mean), and systolic BP (SBP_min) during the ED stay. Consistent presence of these features among all models confirms the importance of persistently abnormal vital signs as a predictor of organ dysfunction and

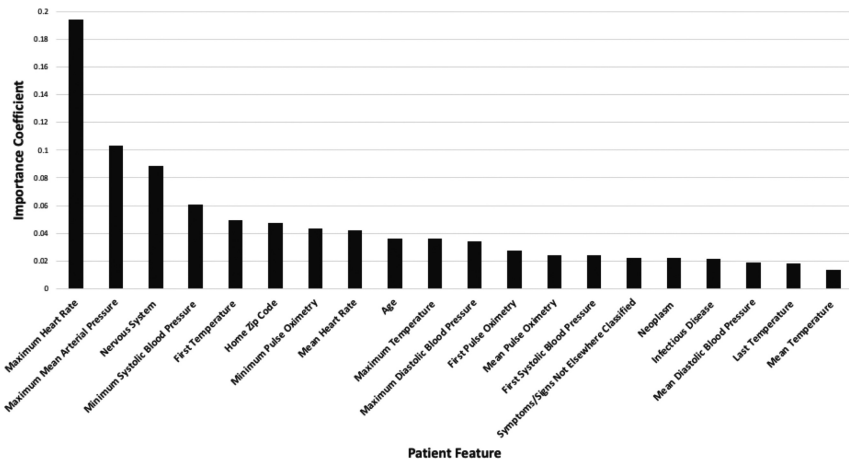

FIGURE 5. Classification and regression tree model: top quartile of patient features predicting sepsis (N = 35,074). DBP indicates diastolic blood pressure; INF, certain infectious and parasitic diseases; NEO, neoplasms; NVS, diseases of nervous system; Oxy, oxygenation level measured by pulse oximetry; Resp, respiratory rate; SBP, systolic blood pressure; SYM, symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified.

therefore a harbinger of sepsis. From a sociodemographic standpoint, top models also identified immunization status, age, and zip code as potential risk factors for sepsis. Subpopulation analyses revealed associations between sepsis/critical sepsis and race, insurance type, and zip code.

Our study population demonstrated an increased prevalence of sepsis/critical sepsis among those with undocumented immunization status (1.25%) or not up-to-date (0.7%) compared with those with up-to-date childhood vaccinations (0.4%). Although not prominent on tests of independent variable association (Fig. 3), immunization status remained a top-quartile feature in the best-performing model (RF) used to identify pediatric sepsis (Fig. 4). Expert review of the sepsis case in this study raised concern that the identified association between immunization status and sepsis may be artificially elevated because of a subpopulation of children who were *unable* to be immunized because of comorbidities, such as cancer. Children undergoing cancer treatment are less likely to be fully immunized because they cannot receive certain vaccines because of immunosuppression, and often do not retain immunity to earlier vaccinations.[21] Stratified analysis by oncologic diagnosis demonstrated that immunization status remained a significant feature only among *non*cancer patients. These results suggest that a cancer diagnosis did *not* artificially inflate the association between underimmunization and sepsis; in fact, the association between underimmunization and sepsis becomes negligible in the oncologic subpopulation. There are many reasons why patients may be behind on immunization: lack of access to care, family belief systems around immunization or medical care, household education, and financial or transportation barriers. Furthermore, recent work by Dunnick et al[22] failed to identify an association between underimmunization and increased risk of bacteremia among children aged 2 to 36 months presenting to a PED, as measured by positive blood cultures. Although invasive bacterial disease is a common first step in developing sepsis, the presence of immunization status in our top-performing model likely represents a more complex marker of baseline health status and barriers to medical care.

Machine learning methods also identified an association between *increasing* age and a diagnosis of sepsis. The prevalence of sepsis was 1.3% among adolescents, as compared with 0.6% in school-aged children and 0.38% in under-5-year-old age groups (Table 2). Age was also among the most predictive features in the RF model (Fig. 4). These data stand in contrast to previous studies identifying infants as the highest risk age group among study populations.[23,24] Although this finding needs further investigation, we postulate that otherwise healthy adolescents are at increased risk of underrecognition of sepsis due to multiple mechanisms. First, patients' developmental status increases their risk of underreporting of symptoms, whereas increasing autonomy and less frequent contact with their primary care provider can lead to delays in presentation. Finally, underrecognition of abnormal vital signs for age by health care providers. To address concerns that perhaps older children may have more health issues predisposing them to sepsis, we examined the relationship between age and patient problem list; descriptive statistics did not demonstrate an association between age and increasing number of documented health problems.

Although not present in the top model, CART and subpopulation analysis identified zip code as a top-quartile predictor of sepsis. This finding supports the recent work of Goodwin et al,[25] in which the authors identify higher incidence and mortality rates of severe sepsis among those living in medically underserved areas. In our study, home zip code was incorporated as an aggregate variable; as such, these results highlight the potential significance of zip codes but do not provide discrete associations between patients' geographic location and their risk of sepsis. Dedicated evaluation of sepsis outcomes by zip code and census block, in combination with other sociodemographic factors, would be required to test this association.

Subpopulation analyses also revealed unexpected associations between insurance status and a sepsis diagnosis. Publicly insured patients, who are typically representative of lower socioeconomic status, are historically at greater risk of being underimmunized, experiencing medical comorbidities, and experiencing worse outcomes when diagnosed with sepsis.[3,26,27] Yet, the prevalence of sepsis was *higher* among privately insured patients (0.9%) as compared with the publicly insured patients (0.5%) in our study population. In addition, unique features such as zip code and the FAC problem list category appeared as significant factors only among privately insured patients. These results suggest that privately insured patients may still experience significant health care barriers based on their copays, zip code, transportation, and income. A 2007 survey of Oregon residents found that access concerns were most common among publicly insured families, but "costs were more often mentioned by families with private insurance. Families made a clear distinction between insurance and access, and having one or both elements did not assure care."[28] A 2016 comparative study of children's access to health care demonstrated that children with all insurance types experienced challenges in access to specialty care and that privately insured children, especially those with special needs, reported significantly greater problems with accessing specialty care, frustration obtaining health care services, and out-of-pocket expenses.[29] These results suggest a more complex relationship between classically defined insurance status, social determinants of health, and sepsis.

As a single-center retrospective observational study, this investigation has several limitations. First, several pieces of information were missing from select patient encounters, including home medications, updated problem list, and immunization status. Our study identified only 13,151 patients (37.4%) with problem list data. This may reflect a combination of most pediatric patients being healthy or incomplete provider documentation at the time of presentation or hospitalization. We were also unable to extract home medication lists for this analysis; patients' medications may have yielded additional insights into underreported chronic medical conditions or underappreciated sepsis risk factors.

Our study also identified 4692 patients (13.4%) without documented immunization information; as such, we were unable to assess the relationship between vaccination and risk of sepsis in this population. Rhode Island reports not only a very high rate of pediatric vaccination—with reports of 91% to 97% complete vaccination of children aged 0 to 2 years[30]—but also hosts a significant population of newly immigrated individuals who may have not had the time or resources to achieve complete vaccination. Prospective study of patients with confirmed vaccination status, as well as incorporation of these missing elements, is needed for further evaluation of this association and could reveal additional modifiable patient risk factors.

Although machine learning methods account for complex relationships between variables and provide internal cross-validation, these models are at risk for overfitting because of our small outcome size. These models are also sensitive to our outcome class imbalance; as such, we reported the AU-PRC, in addition to the generally accepted area under the receiver operating curve, to emphasize model performance limitations. More generally, machine learning models are likely to embody, and potentially amplify, any biases present in the data on which they were trained.[31] In particular, limitations around coding the missing documentation for

structured nursing assessment tools (eg, altered capillary refill, altered mentation) may have contributed to an underestimation of these physical examination findings among septic patients. In addition, study population selection was based in part on the encounter-level diagnosis, which would only be available in a retrospective study. Any prospective application would require a study design independent of this criterion. As such, careful consideration of inclusion and exclusion criteria, as well as cohort construction, is a crucial step in any machine learning study.

With respect to generalizability, predictors identified in a single-center study such as ours should also undergo external validation through application to a broader range of practice locations and populations. Once validated, this information could be used to formulate a more nuanced sepsis screening model for ED providers evaluating large numbers of children with infectious complaints. In particular, these models could alert clinicians to sociodemographic risk factors as well as subtle physiologic state changes during ED visits that heighten suspicion for sepsis in borderline or even well-appearing children.

Finally, recent studies evaluating sepsis documentation highlight inaccuracies in ICD coding as a proxy for actual patient disease because many codes are dependent on appropriate clinical and billing interpretation of patient symptoms.[6] Furthermore, the exclusive use of structured data in this study may not have adequately captured the dynamic, complex nature of the patient state. We attempted to address these limitations through expert review and classification of each pediatric sepsis case (N = 191) in this study, although all 35,047 qualifying patient encounters were not hand reviewed. Further investigation should involve natural-language processing of clinician documentation, which would enable far more sophisticated information extraction. Detailed analysis of the narrative text could yield additional sociodemographic context and subtle signs of illness progression not captured in structured medical information.

Despite improvements in detection and treatment, pediatric sepsis remains a life-threatening condition involving a complex interplay of infection, immunologic, genetic, environmental, and sociodemographic risk factors. Given the proportion of sepsis patients who present to an ED for their initial care, we used machine learning to inclusively examine the relationship between patient characteristics available in the ED and a sepsis diagnosis. Top-performing models not only confirmed well-established predictors of serious illness including initial and ongoing tachycardia and changes in BP, but also identified underappreciated relationships between pediatric sepsis and immunization status, increasing age, and patient demographics.

REFERENCES

1. Ruth A, McCracken CE, Fortenberry JD, et al. Pediatric severe sepsis: Current trends and outcomes from the pediatric health information systems database. *Pediatr Crit Care Med*. 2014;15:828–838.
2. Thompson GC, Macias CG. Recognition and management of sepsis in children: Practice patterns in the emergency department. *J Emerg Med*. 2015;49:391–399.
3. Singhal S, Allen MW, McAnnally JR, et al. National estimates of emergency department visits for pediatric severe sepsis in the United States. *PeerJ*. 2013;1:e79.
4. Balamuth F, Weiss SL, Neuman MI, et al. Pediatric severe sepsis in U.S. Children's hospitals. *Pediatr Crit Care Med*. 2014;15:798–805.
5. Angus DC, Linde-Zwirble WT, Lidicker J, et al. Epidemiology of severe sepsis in the United States: Analysis of incidence, outcome, and associated costs of care. *Crit Care Med*. 2001;29:1303–1310.
6. Balamuth F, Weiss SL, Hall M, et al. Identifying pediatric severe sepsis and septic shock: Accuracy of diagnosis codes. *J Pediatr*. 2015;167:1295–1300.e1294.
7. Brierley J, Carcillo JA, Choong K, et al. Clinical practice parameters for hemodynamic support of pediatric and neonatal septic shock: 2007 update from the American College of Critical Care Medicine. *Crit Care Med*. 2009;37:666–688.
8. Evans IVR, Phillips GS, Alpern ER, et al. Association between the New York Sepsis care mandate and in-hospital mortality for pediatric sepsis. *JAMA*. 2018;320:358–367.
9. Goldstein B, Giroir B, Randolph A. International pediatric sepsis consensus conference: Definitions for sepsis and organ dysfunction in pediatrics. *Pediatr Crit Care Med*. 2007;6:2–8.
10. Mathias B, Mira JC, Larson SD. Pediatric sepsis. *Curr Opin Pediatr*. 2016;28:380–387.
11. Children's Hospital Association. Improving Pediatric Sepsis Outcomes Collaborative. Available at: https://www.childrenshospitals.org/Programs-and-Services/Quality-Improvement-and-Measurement/Collaboratives/Sepsis/Sepsis-Collaborative-Frequently-Asked-Questions. Accessed May 15, 2019.
12. Desautels T, Calvert J, Hoffman J, et al. Prediction of sepsis in the intensive care unit with minimal electronic health record data: A machine learning approach. *JMIR Med Inform*. 2016;4:e28.
13. Balamuth F, Gu H, Keele L, et al. Predictive modeling for organ dysfunction in children with suspected sepsis in the emergency department. *Pediatrics*. 2018;142:127.
14. Nemati S, Holder A, Razmi F, et al. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Crit Care Med*. 2018;46:547–553.
15. Scott HF, Colborn KL, Sevick CJ, et al. Development and validation of a predictive model of the risk of pediatric septic shock using data known at the time of hospital arrival. *J Pediatr*. 2019;217:145–151.e6.
16. Goto T, Camargo CA Jr., Faridi MK, et al. Machine learning-based prediction of clinical outcomes for children during emergency department triage. *JAMA Netw Open*. 2019;2:e186937.
17. Healthcare Cost & Utilization Project. Clinical classifications software (CCS) for ICD-10-PCS (beta version). Available at: www.hcup-us.ahrq.gov/toolssoftware/ccs10/ccs10.jsp. Accessed March 1, 2020.
18. American Heart Association. *Pediatric Advanced Life Support Vol 2020*. Dallas, TX: American Heart Association; 2019.
19. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009. Available at: https://www.python.org/downloads/release/python-373/. Accessed May 1, 2020.
20. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.
21. Fayea NY, Fouda AE, Kandil SM. Immunization status in childhood cancer survivors: A hidden risk which could be prevented. *Pediatr Neonatol*. 2017;58:541–545.
22. Dunnick J, Taft M, Tisherman R, et al. Association of bacteremia with vaccination status in children aged 2 to 36 months. *J Pediatr*. 2021;232:207–213.e202.
23. Watson RS, Carcillo JA, Linde-Zwirble WT, et al. The epidemiology of severe sepsis in children in the United States. *Am J Respir Crit*. 2003;167:695–701.
24. Kawasaki T. Update on pediatric sepsis: a review. *J Intensive Care*. 2017;5:47.
25. Goodwin AJ, Nadig NR, McElligott JT, et al. Where you live matters: The impact of place of residence on severe Sepsis incidence and mortality. *Chest*. 2016;150:829–836.

26. Bohanon FJ, Nunez Lopez O, Adhikari D, et al. Race, income and insurance status affect neonatal sepsis mortality and healthcare resource utilization. *Pediatr Infect Dis J*. 2018;37:e178–e184.
27. Kumar G, Taneja A, Majumdar T, et al. The association of lacking insurance with outcomes of severe sepsis: Retrospective analysis of an administrative database. *Crit Care Med*. 2014;42:583–591.
28. Devoe JE, Baez A, Angier H, et al. Insurance + access not equal to health care: Typology of barriers to health care access for low-income families. *Ann Fam Med*. 2007;5:511–518 (Electronic).
29. Kreider AR, French B, Aysola J, et al. Quality of health insurance coverage and access to care for children in low-income families. *JAMA*. 2016;170:2168–6211 (Electronic).
30. Rhode Island Department of Health. National survey: vaccination coverage rates in Rhode Island. 2018. Available at: https://health.ri.gov/data/vaccination/. Accessed November 01, 2020.
31. Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning. Available at: https://doi.org/10.48550/arXiv.1908.09635. Accessed May 1, 2020.