Machine Learning Early Prediction of Respiratory Syncytial Virus (RSV) in Pediatric Hospitalized Patients

Chak Foon Tso (tsofoon@gmail.com)
Dascena
Carson Lam
Dascena
Jacob Calvert
Dascena
Qingqing Mao
Dascena

Research Article

Keywords: Respiratory syncytial virus, machine learning, pediatric infection

Posted Date: June 8th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1738979/v1

Abstract

Respiratory syncytial virus (RSV) causes millions of infections among children in the US each year and can cause severe disease or death. Infections that are not promptly detected can cause outbreaks that put other hospitalized patients at risk. No tools besides diagnostic testing are available to rapidly and reliably predict RSV infections among hospitalized patients. We conducted a retrospective study from pediatric electronic health record data and built a machine learning model to predict whether a patient will test positive to RSV by nucleic acid amplification test during their stay. Our model demonstrated excellent discrimination with an area under the receiver-operating curve of 0.919, a sensitivity of 0.802, and specificity of 0.876. Our model can help clinicians identify patients who may have RSV infections rapidly and cost-effectively. Successfully integrating this model into routine pediatric inpatient care may assist efforts in patient care and infection control.

1. Introduction

Respiratory syncytial virus (RSV) is the most common lower respiratory tract infection in children; nearly all children have been infected by the time they reach two years of age.[1] RSV causes mild infection in most healthy children, with symptoms often including fever, nasal congestion, and mild cough.[1] However, RSV may also result in severe illness requiring hospitalization. Current estimates suggest that nearly 60,000 children under five years of age are hospitalized with RSV annually in the United States.[2] Hospitalization rates are high among infants < six months old, particularly for those born prematurely,[3] The Centers for Disease Control and Prevention (CDC) estimates that 1–2% of RSV infections in this age group result in hospitalization.[2,4] Other risk factors include premature birth, chronic pulmonary or congenital heart disease, immunodeficiencies, or neuromuscular disorders.[1,2] Pediatric patients hospitalized with RSV may require intensive care unit (ICU) admission and mechanical ventilation, which is associated with substantial healthcare spending both during and following treatment.[5]

RSV can be detected in infected children through polymerase chain reaction (PCR) testing and, less accurately, through rapid antigen testing.[6,7] Because RSV infections are extremely common in children, current guidelines from the American Academy of Pediatrics recommend against routine screening for RSV in young children presenting with respiratory infection,[8] noting that a positive RSV test generally does not change the course of care for patients whose infection can be managed in an outpatient setting. However, RSV testing can be informative for patients treated in hospital settings, where it may help to identify infected patients in need of isolation to prevent outbreaks, as well as identify vulnerable patients in need of additional monitoring and supportive management.[8]

The utility of machine learning algorithms (MLAs) to discriminate between COVID-19 and other viral lower respiratory infections in pediatric patients has been established in previous research.[9] To guide appropriate use of RSV testing, we have developed a MLA to identify pediatric patients who have RSV upon hospital admission. We present a clinically useful MLA that uses individualized demographics and

vital signs data that are routinely collected early upon hospital admission. Infections are substantially enriched among patients identified as high-risk for RSV by our MLA, which demonstrates its utility as a rapid screening tool to help clinicians more efficiently target patients for confirmatory testing and response.

2. Materials & Methods

2.1 Dataset

In this study, a large commercially-available electronic health record (EHR) database was used that collects data from over 700 inpatient and ambulatory care sites located in the United States. Clinical, claims, and other medical administrative data are included in the database. Data was obtained from all emergency department and inpatient encounters for the year 2019. Inclusion criteria included children aged five or younger with at least one measurement of all the required data inputs present within the first two hours of hospital admission. In compliance with the Health Insurance Portability and Accountability Act, all data were de-identified prior to extraction and retrospective analysis. We did not require Institutional Review Board approval, as this project constituted non-human subjects research per 45 Code of Federal Regulations 46.102.

2.2 Data processing

For each patient encounter, only measurements available within the first two hours of admission were used as inputs to predict RSV positivity. The model used the following inputs, which were all required for the model to make a prediction: age, sex, systolic blood pressure (SysABP), diastolic blood pressure (DiasABP), heart rate (HR), respiratory rate (RespRate), body temperature (Temp), peripheral oxygen saturation (SpO2), height, and weight. Time varying features (e.g., clinical measurements) were summarized by the first, mean, and last measurement within the first two hours of admission; those three summary statistics were used as features in the model. An 80/20 train/test split via randomization was used.

2.3 Gold standard

A positive result of RSV nucleic acid amplification tests (NAATs) such as a PCR test, either from a stand-alone test or as part of a respiratory disease panel, was considered the positive label for our model. NAATs are considered the clinical gold standard for diagnosis of RSV.[8,10] All other RSV tests, such as antigen tests, were disregarded by the model. As the model presented here is a binary classification model, all non-positive encounters were automatically considered negative.

2.4 Cohort definition

The attrition process for the MLA is illustrated in Fig. 1. We excluded some patients who had a positive RSV test result, as we could not conclusively determine that it was resulting from an NAAT. We also

excluded patients whose RSV test samples were collected within the first two hours after admission. The final population consisted of 54,413 patients who were randomly split into train and test sets.

2.5 Machine learning nodel

We used XGBoost (XGB, or extreme gradient boost), a class of gradient boosted decision tree implemented it using the XGBoost library in Python.[11,12] We took advantage of the versatility of the algorithm as XGB is highly interpretable and performs well for an imbalanced dataset.[12] A grid search cross-validation was performed to determine the optimal parameters. The parameters used in the final model are reported in Supplementary Table 1.

2.6 Statistical analysis

95% confidence intervals (CIs) were reported for model performance. For the area under the receiver-operating curve (AUROC), bootstrap sampling with replacement of prediction indices was used to generate multiple receiver-operating curves (ROC) and the area under each curve was calculated. We then reported the 5th and 95th percentile values of AUROC. For other performance metrics, 95% CIs were calculated using normal approximation. For the demographics table, Fisher's Exact tests were performed between the positive and negative groups to obtain p-values.

3. Results

To develop and test our models, we used hospital records for 54,413 encounters with patients aged 5 years or younger. Prior to algorithm trigger time, no RSV diagnostic tests had been documented and no RSV tests had been performed within two hours of admission (Fig. 1). These encounters were divided into a training set with 80% (n = 43,530) and a holdout test set with 20% (n = 10,883) of encounters. We observed demographic differences in age between encounters with and without positive RSV tests (Table 1). RSV-positive encounters had higher proportions of patients aged 1 to 3 years ($p < 0.001$); encounters without positive RSV tests had higher proportions of patients who were aged less than one year ($p = 0.002$) or aged 4 to 5 years ($p < 0.001$) or preterm birth ($p < 0.001$). The prevalence of RSV in the holdout test set, as measured by NAAT, was 1.8% (n = 197 RSV-positive), with a test positivity rate of 18.7%.

Table 1
Demographic data of non RSV positive and RSV positive patients with hospital encounters included in the holdout test set.

Demographics	Training set (N = 43,530)	Testing set		
		Non RSV Positive (N = 10,686)	RSV Positive (N = 197)	p-value
Below 1 years old	21,204 (48.7%)	5,244 (49.1%)	61 (31.0%)	0.002
1–3	12,552 (28.8%)	3,024 (28.3%)	120 (60.9%)	p < 0.001
4–5	9,772 (22.4%)	2,418 (22.6%)	16 (8.1%)	p < 0.001
Unknown Age	2 (0.0%)	0 (0.0%)	0 (0.0%)	1
Male	2,363 (54.3%)	5,791 (54.2%)	107 (54.3%)	1
Female	19,743 (45.4%)	4,860 (45.5%)	90 (45.7%)	1
Unknown Sex	153 (0.4%)	35 (0.3%)	0 (0.0%)	1
White	24,065 (55.3%)	5,985 (56.0%)	118 (59.9%)	0.594
Hispanic	5,184 (11.9%)	1,234 (11.5%)	23 (11.7%)	0.911
Black	5,997 (13.8%)	1,441 (13.5%)	21 (10.7%)	0.342
Asian	1,142 (2.6%)	257 (2.4%)	8 (4.1%)	0.158
Other/Unknown	7,142 (16.4%)	1,769 (16.6%)	27 (13.7%)	0.439
Preterm Birth	2,786 (6.4%)	713 (6.7%)	1 (0.5%)	p < 0.001
Smoking Exposure	240 (0.6%)	78 (0.7%)	0 (0.0%)	0.407
Congenital Heart Defects	485 (1.1%)	140 (1.3%)	0 (0.0%)	0.185
Neuromuscular Disorders	7 (0.0%)	1 (0.0%)	0 (0.0%)	1
Down Syndrome	77 (0.2%)	20 (0.2%)	1 (0.5%)	0.320
Cystic Fibrosis	11 (0.0%)	7 (0.1%)	1 (0.5%)	0.137
Chronic Lung Disease	266 (0.6%)	80 (0.7%)	1 (0.5%)	1
Pediatric Immunodeficiency	70 (0.2%)	15 (0.1%)	0 (0.0%)	1
RSV PCR Test Performed	4,175 (9.6%)	851 (8.0%)	197 (100.0%)	p < 0.001
RSV PCR Test Positive	719 (1.7%)	0 (0.0%)	197 (100.0%)	p < 0.001

Figure 2 shows the ROC of the XGBoost model. The AUROC for our model was 0.919, demonstrating exceptionally high accuracy in distinguishing RSV-positive encounters as positive and non-RSV-positive

encounters as non-positive in a binary classification task. Fixing the sensitivity of the model to 0.80 yielded a specificity of 0.876 (Table 2).

Table 2
Summary of algorithm performance metrics.

Performance Metric	Value (95% CI)
AUROC	0.919 (0.906–0.932)
Sensitivity	0.802 (0.746–0.858)
Specificity	0.876 (0.87–0.882)

AUROC: area under the receiver-operating curve (no-skill baseline = 0.50). Optimal specificity was determined with a minimal sensitivity of 0.8.

To determine which features of patient encounters most strongly influenced our model's prediction of RSV, we generated summary plots from Shapley additive explanations (SHAP) analyses (Fig. 3).[13] Our model showed a strong dependence on patient weight and age, together with systolic or diastolic blood pressure and high respiratory rate.[14] These results show that our model's ability to successfully distinguish future RSV positivity among hospitalized pediatric patients is most strongly dependent on vital signs and clinical data that are routinely and rapidly collected at patient point-of-care.

4. Discussion

In this study, we developed an MLA to rapidly and systematically predict a positive RSV NAAT test among hospitalized pediatric patients. This algorithm used inputs that are routinely collected and reported in patients' EHRs within two hours of admission to predict a positive NAAT for RSV later in the same admission. Our work demonstrates the utility of leveraging machine learning techniques to rapidly predict previously unidentified infections among hospitalized patients.

There are two major innovations in our study that substantially contribute to the field. First, our study focuses specifically on identifying likely RSV infections rapidly upon presentation to a hospital emergency room. This differs from previously developed MLAs focused on pediatric RSV infections, which have focused either on predicting future RSV diagnosis, hospitalization, or severe progression of disease in the months to years following data collection, or on identifying RSV infections among pediatric patients that were already hospitalized with known symptoms of respiratory viral infection.[15–17] Several of these previous algorithms also were developed using data only from preterm infants,[16,17] thereby limiting their generalizability as compared to our MLA. As a preventive tool, Heaton *et al.* developed an MLA to predict seasonal RSV outbreaks to allow for timely immunoprophylaxis injections for children predisposed to poor infection outcomes.[18] Other studies using MLAs that predict suitable treatment courses[19] or patient outcomes[15,20] for bronchiolitis patients, a disease commonly caused by RSV, require a proper diagnosis prior to running the algorithm. These RSV preventive and treatment studies do not

address the need for broad screening of incoming pediatric patients and rapid identification of RSV infected patients. Our study therefore provides unprecedented utility among RSV-focused MLAs for hospital healthcare providers to improve the efficiency and accuracy of their initial care for pediatric patients. Second, our MLA is designed to predict RSV positive tests without requiring detailed patient data that require surplus time and effort over standard-of-care protocols performed early in hospitalization. This differs from previously developed risk scores or MLAs that required inputs of ICD diagnosis codes, transcriptome data, and/or documentation of specific symptoms that take additional time to collect and log in patients' EHRs.[16,17,21–23] The relative simplicity of our MLA indicates that integration into hospital settings would be more efficient and immediately useful to clinicians who care for pediatric inpatients.

If successfully implemented as a rapid, preliminary RSV screening system in a hospital setting, our algorithm could provide several primary services to healthcare providers caring for pediatric patients. First, it could be used as a tool for identifying patients to be enrolled or not enrolled in cohort studies or clinical trials that involve active RSV infection - either to include or exclude patients who are actively infected.[24] This would save clinical researchers time and effort by substantially narrowing their scope of viral testing. Second, it could help hospital infection prevention personnel to more quickly identify infected patients who may need to be placed on additional precautions to prevent healthcare-associated transmission of RSV. Outbreaks of RSV in pediatric hospital settings are well documented and have been shown to contribute to increased patient morbidity, mortality, and complexity of care.[6,25,26] Third, our algorithm could better inform delivery of care for infected patients by identifying them more rapidly and with greater efficiency of viral testing. Taken together, these advantages could be leveraged particularly well in tertiary care research and teaching hospitals that would benefit from an efficient alternative to established risk scores or systematic viral testing to identify infected patients.

There are several limitations to this study. First, the use of NAAT testing for RSV as a "gold standard" likely excluded many diagnoses of infection by rapid antigen detection, which may have skewed the RSV prevalence and predictive power of the MLA. Second, we did not include data on the presence or absence of respiratory symptoms that are known to be strong predictors of RSV infection,[2,8,27] because these data were often missing from EHRs of the patients included in this study. Future directions of this research could potentially be improved by considering RSV diagnoses made by rapid antigen testing. Additionally, future studies should include the presence or absence of known symptoms of acute respiratory disease to identify patients with RSV.

5. Conclusions

The model we present in this study performed well in identifying RSV infections among pediatric inpatients at the time they presented to the hospital, using clinical data that are routinely collected in the first two hours following admission. Our model demonstrates utility for clinicians who would benefit from rapidly identifying RSV infections among pediatric inpatients for purposes of infection prevention, clinical trial enrollment, or management of care. Future directions in the field include refining diagnostic

algorithms by including more detailed patient data and the development of new models focused on other infectious diseases of substantial clinical concern.

Declarations

Competing Interests: All authors who have affiliations listed with Dascena (Houston, Texas, U.S.A) are employees or contractors of Dascena.

Ethics

The data acquisition and subsequent study were determined to not constitute human subjects research due to the use of de-identified data. The use of de-identified retrospective data is classified as a non-human subject study and exempt from Institutional Review Board approval, consistent with 45 Code of Federal Regulations 46.102.

References

1. Paes BA, Mitchell I, Banerji A, Lanctôt KL, Langley JM. A decade of respiratory syncytial virus epidemiology and prophylaxis: Translating evidence into everyday clinical practice. Can Respir J J Can Thorac Soc. 2011;18(2):e10-e19.
2. CDC. Learn about RSV in Infants and Young Children. Centers for Disease Control and Prevention. Published December 18, 2020. Accessed March 15, 2021. https://www.cdc.gov/rsv/high-risk/infants-young-children.html
3. Stein RT, Bont LJ, Zar H, et al. Respiratory syncytial virus hospitalization and mortality: Systematic review and meta-analysis. Pediatr Pulmonol. 2017;52(4):556–569. doi:10.1002/ppul.23570
4. Rha B, Curns AT, Lively JY, et al. Respiratory Syncytial Virus-Associated Hospitalizations Among Young Children: 2015–2016. Pediatrics. 2020;146(1). doi:10.1542/peds.2019-3611
5. Amand C, Tong S, Kieffer A, Kyaw MH. Healthcare resource use and economic burden attributable to respiratory syncytial virus in the United States: a claims database analysis. BMC Health Serv Res. 2018;18. doi:10.1186/s12913-018-3066-1
6. Abels S, Nadal D, Stroehle A, Bossart W. Reliable Detection of Respiratory Syncytial Virus Infection in Children for Adequate Hospital Infection Control Management. J Clin Microbiol. 2001;39(9):3135–3139. doi:10.1128/JCM.39.9.3135-3139.2001
7. Allen AJ, Gonzalez-Ciscar A, Lendrem C, et al. Diagnostic and economic evaluation of a point-of-care test for respiratory syncytial virus. ERJ Open Res. 2020;6(3). doi:10.1183/23120541.00018-2020
8. Committee on Infectious Diseases AA of P, Kimberlin DW, Barnett ED, Lynfield R, Sawyer MH. *Red Book: 2021–2024 Report of the Committee on Infectious Diseases*.; 2021. doi:10.1542/9781610025782

9. Nino G, Molto J, Aguilar H, et al. Chest X-ray lung imaging features in pediatric COVID-19 and comparison with viral lower respiratory infections in young children. Pediatr Pulmonol. 2021;56(12):3891–3898. doi:10.1002/ppul.25661
10. Miller JM, Binnicker MJ, Campbell S, et al. A Guide to Utilization of the Microbiology Laboratory for Diagnosis of Infectious Diseases: 2018 Update by the Infectious Diseases Society of America and the American Society for Microbiologya. Clin Infect Dis. 2018;67(6):e1-e94. doi:10.1093/cid/ciy381
11. Python Package Introduction — xgboost 1.4.0-SNAPSHOT documentation. Accessed January 19, 2021. https://xgboost.readthedocs.io/en/latest/python/python_intro.html
12. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM; 2016:785–794. doi:10.1145/2939672.2939785
13. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions.:10.
14. Duarte-Dorado DM, Madero-Orostegui DS, Rodriguez-Martinez CE, Nino G. Validation of a scale to assess the severity of bronchiolitis in a population of hospitalized infants. J Asthma. 2013;50(10):1056–1061. doi:10.3109/02770903.2013.834504
15. Raita Y, Camargo CA, Macias CG, et al. Machine learning-based prediction of acute severity in infants hospitalized for bronchiolitis: a multicenter prospective study. Sci Rep. 2020;10(1):10979. doi:10.1038/s41598-020-67629-8
16. Blanken MO, Koffijberg H, Nibbelke EE, Rovers MM, Bont L, on behalf of the Dutch RSV Neonatal Network. Prospective Validation of a Prognostic Model for Respiratory Syncytial Virus Bronchiolitis in Late Preterm Infants: A Multicenter Birth Cohort Study. Semple MG, ed. *PLoS ONE*. 2013;8(3):e59161. doi:10.1371/journal.pone.0059161
17. Resch B, Bramreiter VS, Kurath-Koller S, Freidl T, Urlesberger B. Respiratory syncytial virus associated hospitalizations in preterm infants of 29 to 32 weeks gestational age using a risk score tool for palivizumab prophylaxis. Eur J Clin Microbiol Infect Dis. 2017;36(6):1057–1062. doi:10.1007/s10096-016-2891-6
18. Heaton MJ, Ingersoll C, Berrett C, Hartman BM, Sloan C. A Bayesian approach to real-time spatiotemporal prediction systems for bronchiolitis. Spat Spatio-Temporal Epidemiol. 2021;38:100434. doi:10.1016/j.sste.2021.100434
19. Mateo J, Rius-Peris JM, Maraña-Pérez AI, Valiente-Armero A, Torres AM. Extreme gradient boosting machine learning method for predicting medical treatment in patients with acute bronchiolitis. Biocybern Biomed Eng. 2021;41(2):792–801. doi:10.1016/j.bbe.2021.04.015
20. Luo G, Stone BL, Nkoy FL, He S, Johnson MD. Predicting Appropriate Hospital Admission of Emergency Department Patients with Bronchiolitis: Secondary Analysis. JMIR Med Inform. 2019;7(1):e12591. doi:10.2196/12591
21. Paes B, Fullarton JR, Rodgers-Gray BS, Carbonell-Estrany X. Adoption in Canada of an international risk scoring tool to predict respiratory syncytial virus hospitalization in moderate-to-late preterm infants. Curr Med Res Opin. 2021;37(7):1149–1153. doi:10.1080/03007995.2021.1911974

22. Mosalli R, Abdul Moez AM, Janish M, Paes B. Value of a risk scoring tool to predict respiratory syncytial virus disease severity and need for hospitalization in term infants: Predicting RSV Hospitalization in Term Infants. J Med Virol. 2015;87(8):1285–1291. doi:10.1002/jmv.24189
23. Jong VL, Ahout IML, van den Ham HJ, et al. Transcriptome assists prognosis of disease severity in respiratory syncytial virus infected infants. Sci Rep. 2016;6(1):36603. doi:10.1038/srep36603
24. Xing Y, Proesmans M. New therapies for acute RSV infections: where are we? Eur J Pediatr. 2019;178(2):131–138. doi:10.1007/s00431-018-03310-7
25. Baier C, Haid S, Beilken A, et al. Molecular characteristics and successful management of a respiratory syncytial virus outbreak among pediatric patients with hemato-oncological disease. Antimicrob Resist Infect Control. 2018;7(1):21. doi:10.1186/s13756-018-0316-2
26. Homaira N, Sheils J, Stelzer-Braid S, et al. Respiratory syncytial virus is present in the neonatal intensive care unit: RSV in NICU. J Med Virol. 2016;88(2):196–201. doi:10.1002/jmv.24325
27. Section on Hospital Medicine AA of P, Gershel JC, Rauch DA. *Caring for the Hospitalized Child: A Handbook of Inpatient Pediatrics*. American Academy of Pediatrics; 2017. Accessed February 22, 2021. http://ebookcentral.proquest.com/lib/beckermed-ebooks/detail.action?docID=5102771

Figures

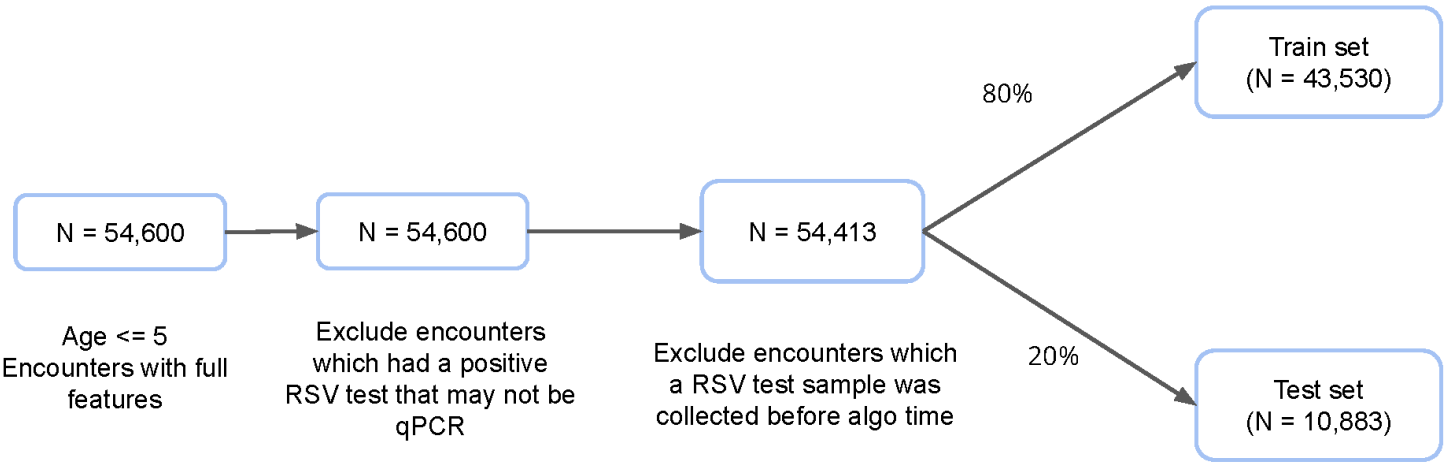

Figure 1

Inclusion criteria for training and testing datasets of patient hospital encounters for algorithm development.

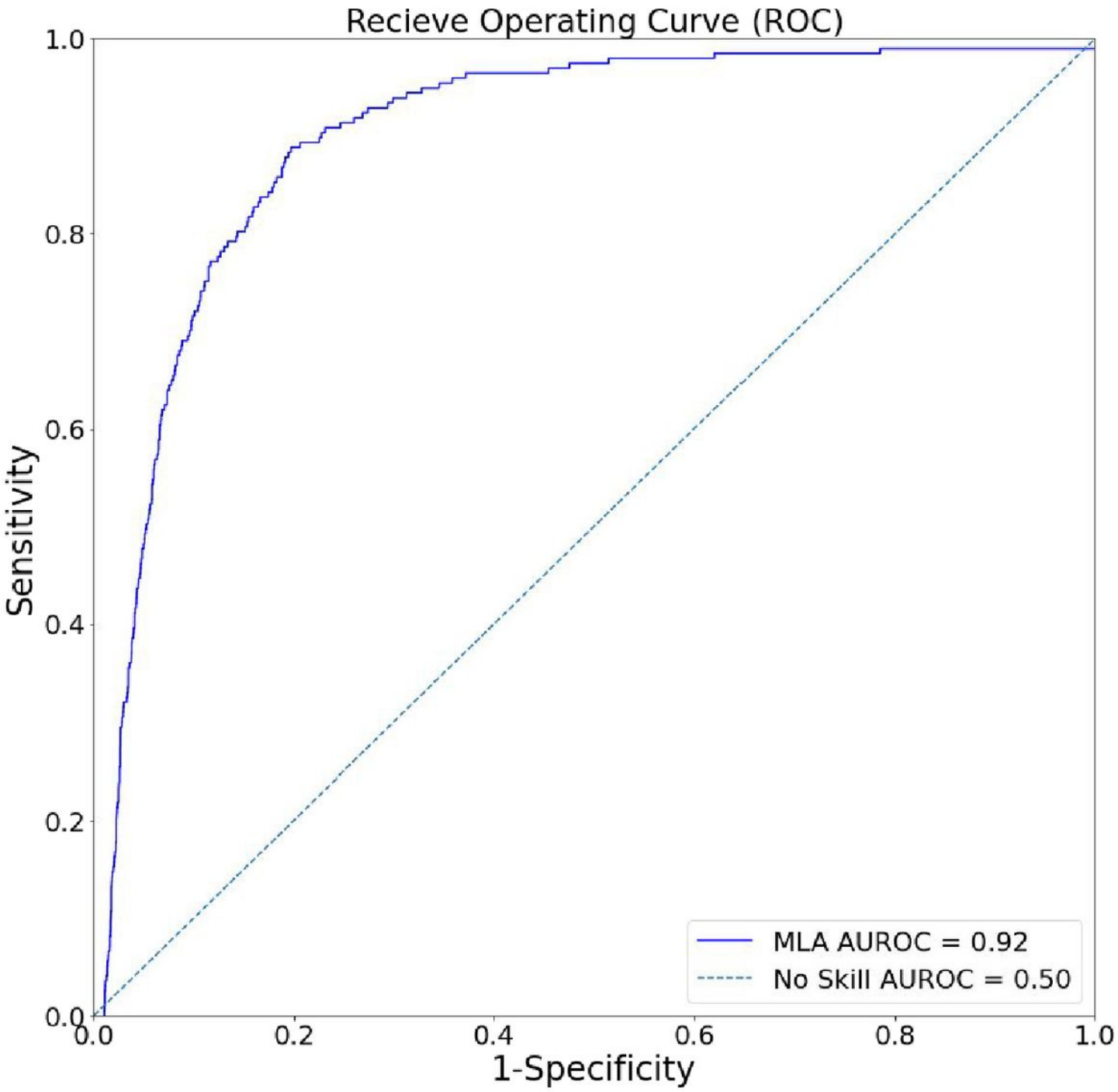

Figure 2

Algorithm discrimination and precision in identifying hospital encounters with future positive RSV tests. The receiver-operating curve (ROC) for the XGBoost model, showing superiority to random chance (gray) in discrimination between RSV-positive and non RSV-positive encounters.

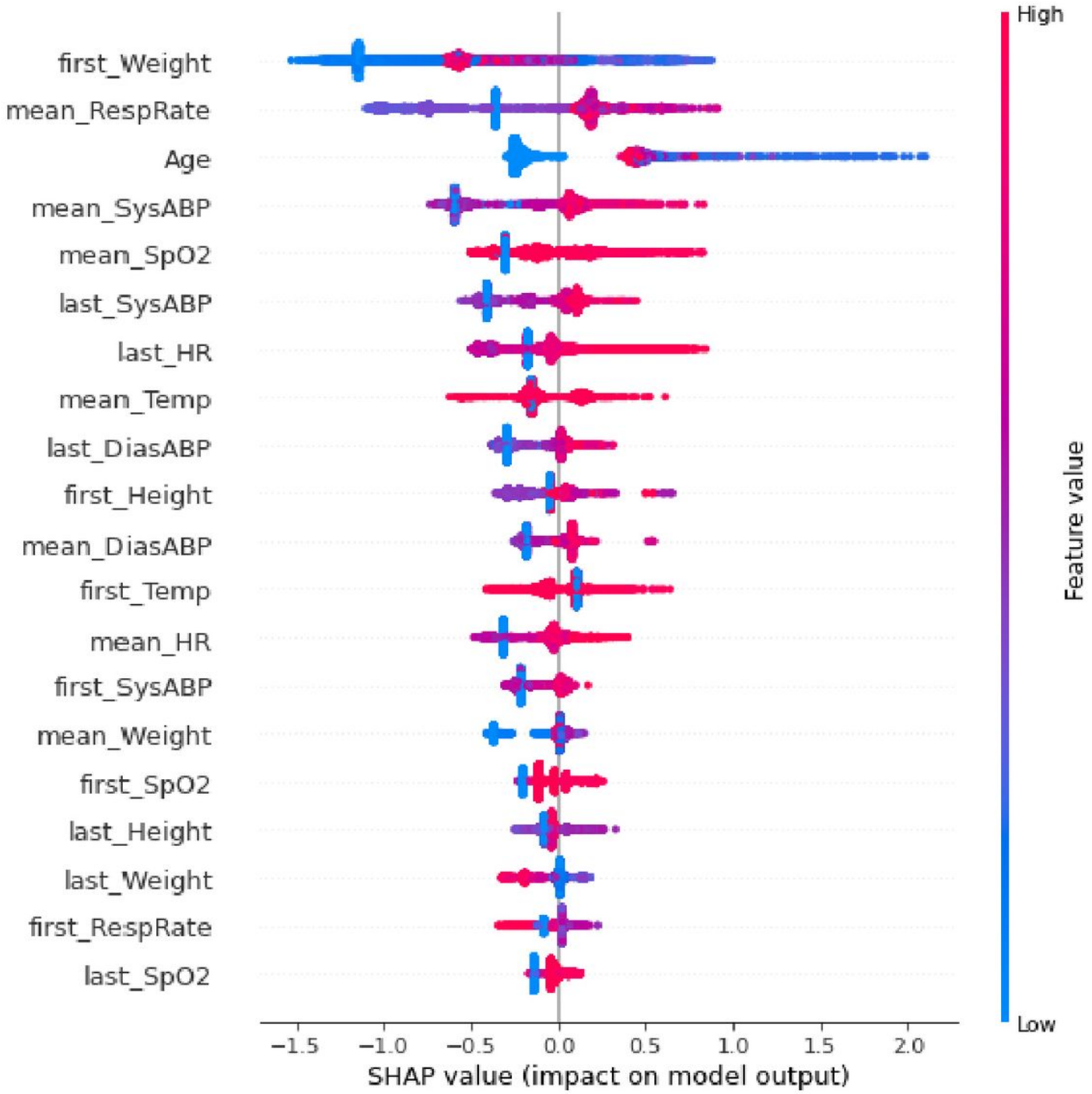

Figure 3

Shapley value plots for degree of model's dependence on specific features. From top to bottom, the relative importance of each feature was ranked. Red dots represent relatively high values of a feature and blue dots represent relatively low values. On the x-axis, the SHAP values (or impact on model output) is plotted. If most of the red dots are on the right of the x-axis, it means high value of that feature (ex. mean DiasABP in this figure) substantially contributes to a positive prediction. SysABP: systolic arterial blood

pressure, DiasABP: diastolic arterial blood pressure, RespRate: respiratory rate, HR: heart rate, SpO2: oxygen saturation, Temp: body temperature.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryMaterials1.docx